12-2015

Evaluating Count Outcomes in Synthesized Single-Case Designs with Multilevel Modeling: A Simulation Study

Kirstie L. Bash
University of Nebraska-Lincoln, bashkirstie@gmail.com

EVALUATING COUNT OUTCOMES IN SYNTHESIZED SINGLE-CASE DESIGNS WITH MULTILEVEL MODELING: A SIMULATION STUDY

by

Kirstie L. Bash

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Arts

Major: Educational Psychology

Under the Supervision of Professor James A. Bovaird

Lincoln, Nebraska

December, 2015

EVALUATING COUNT OUTCOMES IN SYNTHESIZED SINGLE-CASE DESIGNS WITH MULTILEVEL MODELING: A SIMULATION STUDY

Kirstie Lynne Bash, M.A.

University of Nebraska, 2015

Adviser: James A. Bovaird

Complex statistical techniques such as multilevel modeling (MLM) ideally require substantial sample sizes in order to avoid assumption violations. Unfortunately, large between-subjects sample sizes can be impractical and, in some cases, impossible in real-world applications. The use of single-case designs (SCD) allow researchers to overcome this issue. The ability to handle non-normal outcomes appropriately in such single-case designs, however, remains unclear, especially when the outcome reflects recurrent event (count) data.

The purpose of this study is to evaluate the utility of MLM for evaluating recurrent event outcomes in synthesized single-case designs. More specifically, this study seeks to determine the effects of analysis and analytic design decisions when distributional assumptions also vary as a result of the count outcomes. The ability to properly model non-normal distributions in school-based or clinical research settings is critical for two reasons: (1) count data are one of the most prevalent outcomes used in common single-case designs, and (2) it is necessary to avoid biased point estimates and standard errors.

Monte Carlo simulation was used to examine relative bias, mean square error, and confidence interval coverage rates across four simulation conditions: distributional assumption, degree of

freedom methods, sample size, and time-series lengths, where the synthesis of two empirical data sets were utilized to represent the population parameters. As hypothesized, the Negative Binomial distribution performed better, in comparison to the normal distribution and Poisson distribution on relative bias, mean square error, and coverage. The Kenward-Roger and Satterthwaite degree of freedom methods resulted in coverage rates that were closer to the nominal .95 level than the between-within, residual, and containment methods. The results from the sample sizes and time-series lengths were less straightforward than the other conditions.

The results from this study should be used to provide guidance for methodological decisions of synthesized single-case design research. However, researchers should consider their own purpose and research context prior to making methodological decisions, as a single analysis is insufficient for all applied situations.

Acknowledgements

I would first like to thank my graduate advisor, *Dr. Jim Bovaird*, for his assistance with this master's thesis, and for his continuous patience with the question, "do you have a minute?" when it rarely meant just a minute of his time. I would like to thank my "unofficial coadvisor," *Dr. Natalie Koziol*, for her unwavering guidance, as well as her ability to withhold eye rolls during our meetings. I would also like to thank *Dr. Lorey Wheeler* for her excellent feedback at the eleventh hour. Finally, I would like to thank the fellow graduate students who allowed me to hijack their computers. Who knew these analyses would take so long?

I cannot thank my friends and family enough for their support, but more specifically, I would like to thank my partner-in-crime, *Ken*, for his unconditional love, understanding, and patience during this process, and for his uncanny ability to be my personal cheerleader when needed most. Finally, I would like to thank my cat, *Jekyll*, who single handedly prevented countless mental breakdowns at every hour of the day. I forgive you now for stealing my pens.

Table of Contents

List of Figures

Normal Distribution Figures

Poisson Distribution Figures

Negative Binomial Figures

Distribution Comparison Figures

List of Tables

Introduction

Complex statistical techniques such as multilevel modeling (MLM) ideally require substantial sample sizes in order to avoid assumption violations and inaccurate inferences. Unfortunately, large between-subjects sample sizes can be impractical and, in some cases, impossible in real-world applications. For example, researchers in traditional clinical and school-based settings are often interested in interventions to assist students who represent a very small subset of the population, such as children with autism spectrum disorder or other learning disabilities who may cause disruptive behaviors in the classroom and/or in the home environment. In other words, the interest for research conducted in such settings lies in the treatment effect for a small number of participants, and perhaps even just one specific individual. Both limited funding and sample opportunities can restrict researchers from utilizing traditional design and analysis options that might counteract said limitations for data collection within the selective population of interest.

Small *n* studies, including single-case (SCD) and multiple-baseline designs, are common in applied school-based behavioral and clinical research settings within the social sciences and other related fields. Such designs are characterized by few participants but intensive within-subjects data. For the purposes of the current thesis, single-case design refers to the study of one or few individuals, continually assessed over time (Kazdin, 2011, p. 385). The basic feature of SCDs is that participants are repeatedly measured throughout the study, where participants serve as their own "control" prior to the intervention being administered. More specifically, single-case designs typically consist of 20 data points within two phases, baseline and intervention (Shadish & Sullivan, 2011). Recommendations provided by the What Works Clearinghouse (WWC) Standards further suggest that the length of a time series should be no smaller than five

data points in each phase (Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2010).

An emerging body of literature has documented some of the effects of using MLM in these research contexts, including single-case designs (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009), multiple-baseline designs (Shadish, Kyse, & Rindskopf, 2013), and meta-analytic techniques for combining multiple SCDs (Van den Noortage, & Onghena, 2008). Despite recent developments, the ability to handle non-normal outcomes appropriately in SCDs remains unclear, especially when the outcome reflects count, or recurrent event, data.

However, the ability to model recurrent event distributions properly in school-based or clinical research settings is critical for two reasons. First, despite the uncertainty of how to handle the count data most appropriately in given contexts, count data are still one of the most prevalent outcome types present in common SCDs (Shadish et al., 2013). Second, proper modeling of count data is necessary to avoid biased point estimates and standard errors (Shadish et al., 2013). This increased prevalence of count data in SCDs emphasizes the need for methods of preventing and correcting any potential misinterpretations.

The use of SCD (or MBD) data can gain additional momentum when such information is combined across multiple individuals. The focus then shifts from intra-individual effects and the population of an individual's behavior to inter-individual behavior. In this case, the ability to combine, or synthesize, single-case design data allows researchers to gather information about "population" effects. It is often the case that the meta-analysis (i.e., data synthesis) literature focuses vastly more on group comparisons, while excluding single-case design studies (Owens, 2011). The recent surge of synthesized SCD studies (e.g., Owens, 2011; Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012; Van den Noortgate & Onghena, 2008), however,

further emphasizes the growing need and inclination to understand and utilize the methodological approach (i.e., meta-analysis) that was previously underutilized. The synthesis of SCDs (a) strengths the ability to generalize findings beyond study participants, (b) allows for the evaluation of overall treatment effects without losing individual information, and (c) taps into a segment of the literature that was once underdeveloped (Owens, 2011). In short, the synthesis of SCD data can facilitate the increasing growth in the SCD research and literature (Smith, 2012).

The purpose of the present study is to evaluate the utility of multilevel modeling for handling recurrent event (count) outcomes in meta-analysis designs for synthesizing multiple single-case designs using Monte Carlo simulation across various analysis and data generation conditions. More specifically, this study seeks to determine the effects of analysis and analytic design decisions when distributional assumptions also vary as a result of the recurrent event (count) outcomes. In other words, if the data are distributed as "y," what is the impact of the researcher's decision, "*x*," when certain design decisions are made? Design characteristics such as distributional assumptions and degree of freedom methods are often properties that researchers must select appropriately given various sample sizes and time-series lengths. If implemented inappropriately, researchers are often faced with significant impacts to standard errors and statistical inferences.

The long-term goal of this study is to provide recommendations regarding the appropriate methodological decisions/approaches for handling count outcomes in synthesized single-case design data within the same study. The forthcoming recommendations will facilitate improvement towards avoiding the infamous assumption violations in synthesized single-case designs by taking into account the interactions between design decisions and distributional

assumptions of the data. This two-stage research used (1) secondary data obtained through prior behavioral consultation research, and (2) five sets of data simulated through Monte Carlo simulation methods. Obtained parameter estimates in Stage 1 serve as population values in the Stage 2 simulation work. The impact of each distributional assumption (normal, Poisson, and Negative Binomial) with the various study conditions (i.e., sample size, time-series length, and degree of freedom method) were examined. The following section describes the theoretical and empirical background underlying the present study.

Theoretical and Empirical Background

The theoretical and empirical basis for this research is divided into four main sections to discuss the characteristics of: (a) single-case research, (b) meta-analysis, (c) multilevel modeling, and (d) count data. The discussion must first provide an overview of the advantages and issues of single-case design (SCD) in order to transition into the use of multilevel modeling for single-case data. This transition can occur because meta-analytic techniques allow researchers to increase single-case design sample sizes to accommodate the multilevel modeling framework. The discussion ends with a brief description of count data and the various distributional assumptions, which help to address relevant issues with count data.

Single-Case Design Research

The key distinguishing feature of single-case design (SCD) research is the study of one or few individuals (or an aggregate unit such as a classroom or school) characterized by intensive within-subjects data observed over multiple time points for the evaluation of interventions or treatments (Kazdin, 2011). By collecting data successively, one participant (or entity) functions as their own control in a within-subject design (Kratochwill & Levin, 2014; Smith, 2012). SCD research operates under various terms within the same conceptual umbrella, such as single-case, single-subject, $n = 1$, and intra-subject. Although "single-case design" can often imply the study of only one subject, SCDs can also employ more than one individual within the research design (Kazdin, 2011, p. 385). For example, a multiple baseline design across individuals can be implemented to observe behaviors across participants, as such multiple individuals are needed for such a research design. In fact, Shadish and Sullivan (2011) report that, on average, the number of cases/persons per SCD study ranged from 1 to 13 cases, with the average number of cases being 3.64 per study.

SCD research is typically characterized as a *short* time-series approach with significantly fewer data points than 50 – 100 observations that are found in the time-series literature (Shadish & Sullivan, 2011). Single-case design research consists of two experimental phases (baseline and intervention), which average approximately 10 data points or observations per experimental phase (Shadish & Sullivan, 2011). The two phases can be repeated and/or alternated based on the specified design (described below). SCD research highlights the individual variations in the treatment effect, whereas this information is typically lost in between-subject designs that focus on the *average* treatment effect (Barlow, Nock, & Hersen, 2009). More specifically, the nature of the repeated observations within SCDs allow researchers to obtain additional information about the treatment effects of one individual across more than one observational point.

Single-Case Design Variations

There are several variations in the design of single-case research, depending on the specific aims and data availability. The most common variations within psychological sciences are described below, and include the time-series design, the reversal design, the reversal-treatment reintroduction design, the multiple-baseline design, and the changing criterion design.

The most basic SCD design is the time-series design (AB), which exists when a participant is observed several times during the baseline phase and several times during or after the treatment phase (Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012). The baseline phase and the treatment phase are commonly represented as ‘A’ and ‘B,’ respectively. The term “time series” here differs in the SCD context, such that time series in this case are considerably shorter than the 50 – 100 observations for the traditional time series (Greene, 2000). The length of the short SCD “time series” should be no smaller than five data points in each phase; that is, at least five observations in the baseline and at least five observations in the

treatment phase, based on recommendations by the What Works Clearinghouse Standards (WWC; Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2010).

The underlying weakness of the single-case design is the lack of sufficient replications for demonstrating experimental control (Kratochwill & Levin, 2014). In other words, the simple AB design does not allow for a return to baseline a second time to establish that the effect was due to the treatment alone. Therefore, sufficient replications (i.e., no additional baseline phase) is not established for evidence of experimental control. Similarly, the reversal (ABA) design includes a second baseline phase after the initial treatment phase to evaluate if performance returns to baseline. The reversal design, like the basic AB design, fails to meet the minimum criterion for replications that demonstrate experimental control (Kratochwill & Levin, 2014). In order to obtain evidence of experimental control, researchers are encouraged to implement more sophisticated design variations.

The remaining SCDs can be viewed as variations based on the basic time-series design, in which the phases are repeated and/or alternated based on the specified design. First, the reversal-treatment reintroduction (ABAB) design alternates the baseline (A) phases and the treatment (B) phases, where no treatment is administered during the baseline (Kazdin, 2011). For example, an intervention study aimed at improving mathematical scores by increasing the study time through flashcard usage would observe students at baseline before the intervention (A), administer the intervention (B), return to baseline without the intervention (A), then finish with a final intervention phase (B). The underlying purpose of the ABAB design is to demonstrate clearly that any changes on the dependent variable (e.g., study time) are occurring due to the actual treatment alone (e.g., flashcard usage), not due to confounding variables or

threats to validity. This evidence is produced if the researcher observes a return to the baseline levels prior to the final treatment administration.

The multiple-baseline design utilizes multiple baselines with varying time points in the study. Multiple baselines could be characterized as across individuals or across behaviors. For example, the same intervention study that is interested increasing study time could assess the studying habits across three individuals at three separate points in time, in which the initial baseline and intervention phase lengths and starting points varies per individual. The multiple-baseline design across behaviors, on the other hand, allows researchers to observe the behavior changes as an effect of treatment only when the treatment is applied (Kazdin, 2011). Unlike the time-series design or reversal-treatment reintroduction design, the multiple-baseline design has the capacity to demonstrate the experimental control.

Finally, the changing-criterion design demonstrates that the behavior of interest changes gradually during the treatment phase because of the intervention, and not due to other characteristics (Kazdin, 2011). More specifically, the criterion of the dependent variable changes due to the individual matching or exceeding that criterion and advancing to a new level. For example, a student is given praise or a small reward for practicing with mathematical flashcards. In this case, the criterion would be the amount of time spent practicing, in which the criterion is specified prior to practicing as the level that must be met before the reward is given. The criterion will advance to a new level once the student has consistently met the initial criterion established. The changing-criterion design does not require that the intervention be removed or withdrawn to achieve evidence of experimental control, which can be highly sought after depending on the behavior of interest (Kazdin, 2011).

The flexibility of SCDs enables researchers to extend and alter the common design types to best suit their research objectives. The selection of an appropriate research design will be dependent on the intended research purpose, the target sample, and other design characteristics (e.g., treatment withdrawal option, dependent variable). Other SCD variations include simultaneous treatment design, alternating treatments design, mixed designs, and BABA design (administer treatment then baseline). See Kazdin (2011) and Kratochwill and Levin (2014) for more information.

Advantages of Single-Case Designs

The prevalence of single-case design research in the social and behavioral sciences has been rapidly growing over the years (Kratochwill & Levin, 2014). The developments in statistical advancements and increased interest from researchers has contributed to the rise and continual use of SCD data. The availability of limited populations makes SCDs an appealing option for studying those populations when numbers are limited.

The nature of SCD research requires the observation of one or few individuals across several time points. The repeated observations facilitate data collection beyond a single data point to allow for richer, more intensive information from within-subjects that could be unavailable in traditional between-subjects designs. This type of design can establish baseline rates, trends, and variability in the data that are only observable in SCDs (Kazdin, 2011). Single-case designs also ensure that individual variations of the treatment effects can be evaluated—as opposed to the overall treatment effects in group designs. In fact, the most commonly-cited limitation with group designs involves the inability to account for treatment effects at the individual level. The use of single-case designs overcomes this limitation.

The second advantage of SCDs is that, by definition, large sample sizes are not required. It is often the case that researchers are interested in studying individuals that are inherently less prevalent within the population (e.g., autistic children, first-generation college students). While it is possible that "large" samples could be obtained for a specific target population, increasing the sample size may require the use of unavailable funds and/or additional time. The advantage of implementing a SCD is in the ability to study these individuals without the sample size restriction that is common for between-group designs (Van den Noortgate & Onghena, 2003).

Finally, SCDs allow researchers to tailor their research design based on established design variations to be most appropriate for their research purposes. Additionally, the flexible designs allow researchers to study the effects of interventions over more than one time point to assess relevant and significant changes (Owens, 2011). Any concerns about establishing and demonstrating experimental control in a SCD can be diminished by including additional baseline and treatment phases for assessing the impact on the dependent variable. Furthermore, SCD researchers can make methodological changes throughout the experiment, instead of reluctantly completing the entire time series before phase alterations (Kazdin, 2011). It should be noted, however, that these methodological changes should be made only when protocol for such changes has been developed before the experiment begins.

Relevant Issues of Single-Case Designs

The unique aspects that SCDs contribute to the social and behavioral sciences simultaneously elicit certain methodological and statistical issues for data collection and data analysis. The first issue that is frequently discussed in the SCD context is the inability or difficulty with generalizing empirical research findings to the population, when compared to the generalizability of between-subjects research (Kazdin, 2011). The low prevalence and the rather

exclusive participant characteristics creates issues with generalizing findings to individuals outside the study and in less restrictive samples. Further complications can arise when researchers attempt to generalize SCD findings to the between-subjects designs. Additional information on this topic can be found in Van den Noortgate and Onghena (2008).

The other relevant issue for SCDs is the occurrence of repeated observations that create serial dependence due to consecutive observations. This serial dependence is referred to as autocorrelation, in which one time point is correlated with both the preceding and the future time points (Jenson, Clark, Kircher, & Kristjansson, 2007; Kazdin, 2011). Autocorrelation occurs when research designs implement multiple measurement occasions across more than one individual, such that observations within a person will be more correlated than observations between persons. The independence assumption can further be violated when observations are nested within a higher-order unit, and lower level units are dependent on the structural component at the higher level (Snijders & Bosker, 2012). For example, observations of students in one school are likely to be more correlated with one another than with the observations of students in a different school due to the shared environment (Peugh, 2010). Additionally, the higher-level unit (e.g., teacher, school) can have a significant impact on variability within the lower-level units (e.g., students, children); this requires statistical methods that can accommodate for the hierarchical structure accordingly.

Meta-Analysis of Single-Case Design Research

Meta-analysis refers to the quantitative integration of research findings through statistical analysis based on the results from multiple studies (Glass, 1976; Ugille et al., 2012). Individual SCD data can be synthesized across different studies and within the same study through meta-analytic techniques (e.g., Van den Noortgate & Onghena, 2003; Ugille et al., 2012), especially in

cases where between-subjects samples are impractical to obtain for statistical analyses. For example, a multiple baseline design across individuals can synthesize data from within the same experimental study. More specifically, consider a multiple baseline design across individuals in which five children are observed in order to evaluate any changes in disruptive behaviors before and after intervention. The within-subject information as well as the between-subject information are available after synthesis of the single-case design data. The ability to synthesize single-case data provides researchers with the opportunity to examine and evaluate substantive research questions that were once limited by small samples and complex statistical techniques. The empirical basis for synthesizing data using meta-analytic techniques is well documented in the literature, and these techniques have allowed for additional flexibility and for gathering unique sources of new information (Ferron et al., 2009; Van den Noortgate & Onghena, 2003; Ugille et al., 2012). More specifically, the synthesization of SCD data allows for the estimation and testing of both group and individual parameters, which (a) no longer limits the finding to either group or individual, and (b) increases the generalizability of the research outcomes.

General Procedures of Research Synthesis

The general procedures for synthesizing research involve five stages, as outlined in Cooper (1998). Transitioning from one stage to another is considered to be a fluid process that permits the researcher to skip the stage completely, advance into the next stage, or return to the previous stage as necessary. The stages of research synthesis include: (1) problem formulation, (2) data collection or literature review, (3) data evaluation, (4) analysis and interpretation, and (5) presentation of results (Cooper, 1998). Each stage requires that the researcher be responsible for sufficient documentation of progress in order to disseminate the information for future replications.

Briefly outlined, the *problem formulation* stage involves the operationalization of key concepts and variables in the study, as well as distinguishing between studies that operationally defined the concepts differently (Cooper, 1998). This stage in the research synthesis process provides the foundation for the underlying purpose of the study. The *literature review* or data collection stage examines the relevant studies and target populations based on the research purpose and specific criteria of interest. Researchers operating in this stage collect information about the target population and collect data to be used in the subsequent stages. The *data evaluation* stage addresses two components of the literature: (a) the quality differences of the studies, and (b) separates the relevant studies from irrelevant studies (Cooper, 1998). It is important that data evaluation be thorough enough to exclude irrelevant studies, but comprehensive enough to maintain relevant studies.

The *analysis and interpretation* stage incorporates the data points across studies into a synthesized, unified "statement of the problem" (Cooper, 1998). The analysis part of this stage facilitates the use of quantitative procedures, otherwise known as meta-analysis (Glass, 1976); an important extension, because it further increases the replicability and validity of research conclusions. The analysis part can include analyzing the raw data or the effect sizes within separate empirical studies (discussed below). The interpretation part of the fourth stage requires that the research provide evidence based on the analysis part of the inferences produced. The *presentation* stage allows the researcher to present results to describe and disseminate the synthesis process, research findings, and relevant limitations (Cooper, 1998).

Raw Data versus Effect Size Measures

The analysis stage of research synthesization requires the use of raw data or effect sizes obtained from several research studies. These data types are currently represented in two

synthesization approaches for combining SCD data, otherwise known as individual participant data (IPD; Cooper & Pattall, 2009) and aggregate data (AD; i.e., effect sizes).

The individual participant data approach synthesizes raw data that are collected from graphs or tables presented either within several research articles or a single study (Van den Noortgate & Onghena, 2008). If the dependent variable was measured on different scales across the different studies, then the raw data should be standardized before analyses are conducted; this standardization allows for meaningful comparisons. The raw data approach for synthesization allows the researcher to confirm that no errors exist within the data itself (Owens, 2011). More specifically, raw data can be evaluated for errors, such as data input, data manipulation, or incorrect inferences drawn. The scores (or data points) from repeated observations of participants can additionally be grouped together according to the participant they belong to (Kazdin, 2011; Van den Noortgate & Onghena, 2008). Furthermore, the use of raw data creates an easier way to perform complex analyses, such as multilevel modeling, which can address autocorrelation issues in SCD research (Owens, 2011; Van den Noortgate & Onghena, 2008).

The aggregate data approach utilizes effect size measures if raw data are unavailable for analysis. The advantage of the aggregate data approach is that effect sizes are unaffected by the size of the sample when determining small, medium, or large effects for highlighting group differences. Combining effect size measures can also be completed quicker through meta-analysis and usually more cost and time effective (Owens, 2011). Similar to the IPD approach, effect size measures may require standardization, if the dependent variable(s) has not been measured on the same scale, resulting in different effect size outcomes (Van den Noortgate & Onghena, 2008). Mathematically, this standardization can be expressed as (Van den Noortgate & Onghena, 2008):

$$\delta_{jk} = \frac{\beta_{1jk}}{\sigma} = \frac{\mu_{Treatment} - \mu_{Control}}{\sigma} \qquad (1)$$

where Equation 1 refers to the mean difference between baseline phase ($\mu_{Control}$) and treatment phase ($\mu_{Treatment}$) is divided by the within-condition variance (σ).

Simulation results from Ugille et al. (2012) indicated two relevant findings: (1) unstandardized effect sizes performed well under the multilevel meta-analysis approach, and (2) standardized effect sizes also performed well, in limited instances where (a) 30 or more studies were combined, (b) 20 or more measurement occasions per subject were included, and (c) there was homogeneity across/within the studies. The general rule is that effect size standardization is not required for multilevel meta-analysis when participants from the same studies are combined. Small issues arise, however, when standardized samples are somewhat homogenous and/or measurement occasions exceed 20 or more per participant. The aggregate data method fails to work properly when participants are measured over a few occasions.

The decision to combine SCD data through the IAD approach or the AD approach using either raw data or effect size measures, respectively, depends on the data and other information that is available to researchers. For SCD meta-analytic purposes, raw data, effect sizes, standard deviations, and means are typically presented in published empirical studies. SCD researchers have historically relied on visual inspection of raw data and effect size measures, rather than statistical techniques. The recent trend, however, has begun to shift more towards reporting effect sizes more readily. Both meta-analytic approaches (IAD and AD) enable SCD studies to be treated similarly to group-comparison studies and to be analyzed by complex statistical techniques.

Advantages of Synthesizing Single-Case Design Research

Meta-analysis is a highly structured, systematic technique for quantitatively synthesizing findings, such that it requires researchers to produce well-documented accounts of all processes when conducting a meta-analysis (Owens, 2011). These procedures for meta-analysis that must be documented explicitly facilitate the communication and confirmation of replicating research findings within a particular context. Additional replication studies within the behavioral sciences ensures that statements of generalizability to the population are empirically established.

More specifically related to SCD research, meta-analysis becomes a necessary and essential technique for obtaining information from particularly small populations of interest (Van den Noortgate & Onghena, 2008). The increased prevalence of SCD studies calls for methods and techniques that allow for complex analyses, in which researchers can maintain both individual and group level information without the traditional large sample requirements (Owens, 2011). Furthermore, the flexibility within meta-analysis procedures ensures that the methods are easily adaptable for specific research interests and data characteristics.

Meta-analysis techniques also allow researchers to differentiate between overall treatment effects and characteristics that influence the treatment effect (e.g., persons, settings) in single-case designs. Once data have been synthesized by meta-analysis, the next step is to consider the appropriate methods by which to analyze data. The next section discusses the analysis options, specifically multilevel modeling, for synthesized SCD data.

Analysis Options for Synthesized Single-Case Design Data

There are several options for the analysis of SCD data, which vary based on the type and degree of information that the researcher is seeking. *Visual inspection* or visual analysis is the traditional approach for analyzing general SCD data. Visual analysis refers to the graphing of

data such that the data, trend(s), variability, and any overlap can be visually examined to assess the intervention effects (Kratochwill & Levin, 2014). The subjective nature of interpretation for the outcome criteria and decision criteria is the primary issue for visual analysis, where analysis interpretation relies on the researchers' judgments and perceptions. The direction that SCD research has taken is to implement statistical analyses of most SCD research to supplement the interpretations of visual analyses (Parker & Brossart, 2003).

The second analysis option is known as the *percentage of non-overlapping data* (PND), and it determines the treatment effectiveness of synthesized studies by dividing the number of data points that exceed (or overlap) the "highest" data point during baseline by the total number of data points then multiplying the value by 100 (Campbell, 2004; Scruggs, Mastropieri, & Castro, 1987). The PND approach entails three primary limitations in which the treatment effects can be misrepresented when: (1) trends exist in the data, (2) outliers exist in the treatment phase, and (3) treatment produces a negative effect on the outcome (Allison & Gorman, 1993). As such, visual inspection and other inferential statistics are limited in their applications and statistical inferences.

The use of multilevel modeling can also be a desired statistical method for addressing limitations within the synthesized single-case design framework, such as the ability to account for autocorrelation and to account for the recurrent event (count) outcomes commonly present (e.g., Owens, 2011; Rindskopf & Ferron, 2014; Van den Noortgate & Onghena, 2008). The following section outlines multilevel modeling, including the general framework and the single-case design research framework.

Multilevel Modeling in Single-Case Design Research

Multilevel modeling (MLM) falls under several names across the sciences, which include hierarchical linear modeling (HLM), random coefficients modeling, and mixed effects modeling, and has been rapidly increasing in its widespread utility and software advances in recent years. This increase in utility and software has also had a positive impact on the statistical analyses of SCD data. The key distinction between single-level modeling and multilevel modeling is that multilevel modeling can statistically account for the hierarchical structure of empirical data, whereas single-level modeling does not (Heck, Thomas, & Tabata, 2012).

The higher-order levels result in data that are hierarchically-structured, which require MLM analysis to avoid misinterpretations and incorrect estimates. Social and behavioral science data can be classified as hierarchically structured in nature. For example, students nested within teachers, teachers nested within schools, students nested within counselors, etc., in which individuals interact with their social context and therefore, are influenced by the higher group dynamics (Maas & Hox, 1999). Conversely, possible score dependences in single-case designs are taken into account when the hierarchical structure is modeled.

The general multilevel modeling framework is conceptually defined by the hierarchical system of regression equations (Hox, 1998), in which there is a regression equation for each hierarchical level, based on the group-level coefficients. A simple linear model can be expressed in Equation 2 as,

$$y_{ij} = \beta_{0j} + x_{ij}\beta_{1j} + \varepsilon_{ij} \qquad (2)$$

where y_{ij}represents the outcome variable at group level (j) for person (i), $\beta_{0j}$is the group-specific (j) intercept (see Equation 3), $x_{ij}\beta_{1j}$is the systematic component (see Equation 4), and $\varepsilon_{ij}$is the unexplained variance. The group-specific intercept (β_{0j}) is assumed to be normally

distributed with a mean (β_0) and standard deviation (σ_{u0}), whereas the overall intercept (β_0) is considered to be the fixed effect, and the difference ($u_{0j} = \beta_{0j} - \beta_0$) is the random effect (Gill & Womack, 2013). The group-level regression model can then defined in Equations 3 and 4 as,

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} , \quad (3)$$

and

$$\beta_{1j} = \gamma_{11} Z_j + u_{1j}. \quad (4)$$

where the $[\gamma_{00} + \gamma_{01} Z_j]$ and $[\gamma_{11} Z_j]$ represent the level-two fixed coefficients, and the residuals at the group-level (u_{0j}, u_{1j}) are assumed to be multivariate normally distributed (Maas & Hox, 2005).

The utility of multilevel modeling, specifically for synthesized SCD research, allows for obtaining information about the intervention effects during the treatment phase (Baek et al., 2014); for example, the degree of treatment effects across participants and across studies can also be examined using this statistical approach (Rindskopf & Ferron, 2014). Multilevel models for synthesized SCDs are similar to the general framework, and accounts for the within-person variation, as demonstrated by the regression equation (Van den Noortgate & Onghena, 2008b):

$$y_{ijk} = \pi_{0jk} + \pi_{1jk}(condition) + \varepsilon_{ijk}, \quad (5)$$

where Equation 5 includes the outcome variable (y_{ijk}) in regards to person *j* on occasion *i* for study *k*. The "condition" variable represents a dummy-coded variable which specifies the measurement phase during baseline (0) and treatment (1). Subjects are further represented by two scores; the π_{0jk} parameter represents the expected outcome (score) during the baseline phase, whereas the π_{1jk} represents the expected outcome (score) during the treatment phase. Finally, the within-condition error variance is expressed as ε_{ijk}. Therefore, the second-level

regression equations, which model the across-participant variation, are expressed in Equations 6 and 7 as:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \quad (6)$$

and

$$\pi_{1jk} = \beta_{10k} + r_{1jk} \quad (7)$$

where β_{00k} represents the average baseline level (fixed effects) and β_{10k} represents the average treatment effect for study *k*, with error terms (r_{1jk} and r_{0jk}) to represent the random deviation.

A third-level model can also be specified to account for cross-study variation or across another higher-level unit such as teacher or organization. For third-level regression equations that model the across-participants variation, this can be expressed in Equations 8 and 9 as:

$$\beta_{00k} = \gamma_{000} + u_{00k} \quad (8)$$

and

$$\beta_{10k} = \gamma_{100} + u_{10k}, \quad (9)$$

where γ_{000} (fixed effects) represents the average base or grand mean baseline level and γ_{100} represents the average effect (i.e., grand mean difference between baseline and treatment phases). The random effects for the second level are represented as u_{0j} and u_{1j}, respectively, and allow for variation in baseline levels and treatment effects among participants (Owens, 2011). The fixed effects ($\gamma_{000}\ and\ \gamma_{100}$), and the variance and covariance parameters ($\sigma_c^2,\ \sigma_{u0}^2,\ \sigma_{u1}^2,\ \sigma_{u0u1}^2$) are the specific model parameters of interest (Van den Noortgate & Onghena, 2003a). Because the three-level multilevel model allows for variation within participants, variation between participants of the same study, and variation between participants of different studies, the dependency is accounted for among observations from the three levels (Van den Noortgate & Onghena, 2008).

In sum, multilevel modeling for single-case designs is comprised of a linear model at each hierarchical level, which describes the variation of scores at the within-group level (level 1). In a simple AB design, an individual's change in intervention over time is classified in the level-1 model. More specifically, the treatment effect for one individual in the level-1 model is measured by the difference between the baseline mean and the treatment mean (β_0), which is considered the most important parameter in this case (Jenson et al., 2007). The first-level regression equation must be specified and the second-level coefficients must be allowed to vary in order to combine non-continuous data over cases (Baek et al., 2014). At the second level, the regression coefficients of the first level are allowed to vary, and represent the variation across participants. The third level models the variation *across* studies that were gathered during the meta-analysis process.

Advantages of Multilevel Modeling in Synthesized Single-Case Design Research

The advantages of using MLM in synthesized SCDs are well documented in the literature. For instance, MLM can be used to account for any hierarchical structures present in the social and behavioral sciences data (Gill & Womack, 2013), specifically in the context of synthesized single-case designs where designs may include children nested within teachers and schools. Multilevel modeling can also be used to account for the interdependency that results from the successive, repeated observations of a single-case design that are nested within higher level units (Baek et al., 2014). Because multilevel modeling accounts for observation dependency, the independence assumption violation, as well as the resulting Type I errors and biased parameter estimates can be minimized (Hox & van de Schoot, 2013; Peugh, 2010).

The advancements in sophisticated software options for complex data structures have facilitated the widespread interest and utility of using multilevel modeling, including the

specification of generalized linear multilevel models with non-normal outcomes, non-nested hierarchies, longitudinal design considerations, etc. (Gill & Womack, 2013). Additionally, multilevel modeling can make parameter estimation more efficient, even in instances where fewer number of scores per unit are present (Bryk & Raudenbush, 1992). The benefit of multilevel modeling, especially within synthesized SCDs, is the flexibility to adapt and extend the model further according to the desired specificity of the SCD data (Van den Noortgate & Onghena, 2003a). More specifically, research elements such as data type (e.g., count, frequencies, or proportions), autocorrelation, and linear or nonlinear trends can be accounted for and modeled appropriately in multilevel modeling (Rindskopf & Ferron, 2014).

Considerations for Single-Case Design Research within the Multilevel Modeling Context

Single-case design data analyzed within a multilevel modeling framework violates a number of statistical assumptions. The first issue is related to sample size. Multilevel modeling, specifically the asymptotic nature of the maximum likelihood (ML) estimator, requires that the sample size be sufficiently large. While "large" is a relative approximation, the minimum sample size recommendation based on simulation work is approximately 50-60 participants (level-2 units) for smaller models (Eliason, 1993; Hox & van de Schoot, 2013). This sample-size requirement in MLM is predominantly of concern in the highest level because the smallest sample sizes correspond to the highest hierarchical level in the model (Hox & van de Schoot, 2013). The inherent nature of SCD data, where samples are substantially smaller than between-subject designs, reduces the likelihood of obtaining the recommended sample size; in fact, the average SCD study retains 3.64 cases per study for single-level analyses (Shadish & Sullivan, 2011). To address this small sample size issue, three primary options can be implemented: (1) increase the sample size, which can be impractical in most SCD circumstances, (2) implement

meta-analytic techniques to synthesize data across or within studies, or (3) increase the length of the time series.

Second, autocorrelation can also have two potential impacts on Type I and Type II error rates. That is, a significant positive autocorrelation causes an overestimation of Type I errors, and results in a statistically significant effect when one is not actually present (Jenson et al., 2007). Conversely, a significant negative autocorrelation causes an underestimation of the Type I errors. A negative autocorrelation causes a non-statistically significant effect when one, in fact, exists (Jenson et al., 2007). Concerns about balancing Type I and Type II error rates have been debated, in which Baer (1977) argues that single-case design researchers favor "very low probabilities of Type I errors, and correspondingly high probabilities of Type 2 errors" (p. 167); that is, they are willing to overlook potentially effective treatments with small effects because they are interested in detecting powerful treatment effects. Third, multilevel modeling relies on the multivariate normality assumption (MVN), in which residuals are normally distributed, without skewness or kurtosis. The MVN assumption is rarely met in single-case design data, given that count data is highly prevalent in SCDs. Response (or dependent) variables that lack a normal distribution results in incorrect asymptotic standard errors, as well as inaccurate significant tests and confidence intervals (Hox & van de Schoot, 2013).

The considerations presented here highlight the need of multilevel modeling for synthesized single-case designs, where meta-analysis can aid in increasing sample size and where multilevel modeling can account for the autocorrelation, as well as the non-normal outcomes, within single-case design data. The use of meta-analysis techniques to increase sample size allows single-case design researchers to implement statistical techniques, such as multilevel modeling.

Meta-Analytic Multilevel Modeling for Single-Case Design Research

Van den Noortgate and Onghena (2003a, 2003b, 2008a, and 2008b) have made significant contributions towards the advancement of using meta-analytic multilevel modeling in the context of single-case design (SCD) research. Van den Noortgate and Onghena (2003a) first demonstrated the process of combining SCDs across studies through multilevel modeling, as the Busk and Serlin (1992) approach was compared against the MLM approach. The Busk and Serlin (1992) approach actually consists of three approaches that differ in the assumptions for obtaining effect size estimates (i.e., no-assumption, equality of variances, and normal distribution, respectively). Van den Noortgate and Onghena (2003a) focused specifically on the third approach, which specifies that effect size measures are produced based on the pooled within-phase variance for each subject. The third approach also establishes the assumption of normality and equal within-phase variances. Van den Noortgate and Onghena (2003a) selected this approach in particular, because it includes the "possibility of estimating and testing the individual and overall effect sizes." Their findings indicated that the MLM approach provided more advantages, including flexibility, over the Busk and Serlin (1992) approach. Furthermore, unlike the MLM approach, the Busk and Serlin (1992) approach cannot separate the sampling variation and the "true" between-cases variation.

A continuation study by Van den Noortgate and Onghena (2008a) further assessed the use of multilevel modeling for the meta-analysis synthesis of SCD data. The focus was specifically on empirical examples for combining data based on the raw data versus the effect size measures, in addition to comparing effect sizes across single-subject and group-comparison studies. For single-case and between-subjects synthesis, the effect sizes can be represented in Equations 10 and 11, respectively, as:

$$\delta_{SS} = \frac{\mu_B - \mu_A}{\sigma_{within\ phase}} \qquad (10)$$

and

$$\delta_{GC} = \frac{\mu_E - \mu_C}{\sigma_{within\ group}}, \quad (11)$$

where the difference between condition means ($\mu_B - \mu_A$) is divided by the within-condition standard deviation ($\sigma_{within\ phase}$). This is similar to the between-subjects design such that the equation differs only in that the first component is now the difference *in* condition means; conceptually, however, effect sizes in SCDs use the scores from the same participant, whereas the between-subjects designs use scores from different, independent subjects (Van den Noortgate & Onghena, 2008a). The effect size measures between SCD and between-subjects designs must be comparable for synthesization across the designs (Van den Noortgate & Onghena, 2008a).

The findings suggested that combining single-subject and group-comparison effect sizes can yield promising results and more reliable estimates of the unknown parameters. Because this synthesis of group-comparison and single-subject is still relatively new, it is suggested that researchers report the meta-analyses from both group-comparison and single-subject studies alongside the overall meta-analysis that combined these two distinct research designs (Van den Noortgate & Onghena, 2008a). This study also highlighted the issue with data that fail to conform to the independently normally distributed assumption for future research.

Similarly, Ferron, Bell, Hess, Rendina-Gobioff, and Hibbard (2009) examined the multilevel modeling of multiple-baseline data across five simulation conditions: sample size, baseline-level variance, treatment-effects variance, repeated observations, and autocorrelation in Level-1 errors. The five different degree of freedom methods (i.e., residual, containment, between-within, Satterthwaite, and Kenward-Roger) were also examined for capturing the average treatment effect. The first finding was that coverage rates decreased across the methods

of estimating degree of freedom as the sample size increased (Ferron et al., 2009). More specifically, as the number of participants increased, the containment method showed less overcoverage, the residual and between-within methods showed less undercoverage, while the Satterthwaite and Kenward-Roger methods showed coverage rates close to the average nominal level (.95) when autocorrelation was modeled.

Second, coverage rates showed a minor decrease across all five degree of freedom methods as the time series length increased when autocorrelation was modeled (Ferron et al., 2009). More specifically, the average coverage decreased when the time series length increased from length of 10 (.955) to a length of 30 (.950) for the Kenward-Roger method. The Satterthwaite method showed a similar decrease in coverage from a series length of 10 (.949) to a series length of 30 (.948). Ferron et al. (2009) suggested the use of either the Satterthwaite or Kenward-Roger methods over the other three methods, further indicating that the undercoverage of the residual and between-within methods lead to the recommendation of avoiding these methods. Ferron et al. (2009) provided insight into the interactions between sample size, repeated observations, and autocorrelation in a SCD study using multilevel modeling. Ferron and colleagues (2009) did not utilize meta-analytic methods for combining SCD data, such that the simulated sample sizes (i.e., 4, 6, and 8 subjects) were relatively small in comparison to other meta-analysis and large between-subjects samples.

Furthermore, their research assumed a normal distribution of the data, which rarely occurs within SCD data (Shadish & Sullivan, 2011; Smith, 2012). Count data are one of the most prevalent outcomes used in common SCDs (Shadish et al., 2013), and can assume numerous distributions (e.g., Poisson distribution, Negative Binomial distribution, etc.). When

distributional assumptions are assumed to be normal when the data are, in fact, non-normal, the results are likely to include incorrect point estimates and standard errors (Shadish et al., 2013).

Implications for Count Data in Single-Case Design Research

Previous research has made significant strides towards better understanding the use of meta-analysis procedures for SCD research and the utility of such for multilevel modeling. However, the limitation of previous SCD research lies in meeting the assumption of a normal distribution assumption for data analysis. For some SCD data, this assumption holds true; however, most single-case design data are count data, or recurrent event data, that are rarely normally distributed in real-world applications (Shadish et al., 2013). Examples of count data include the number of the interactions initiated by students with disabilities (Shadish et al., 2013), the number of physician visits (Pohlmeier & Ulrich, 1995), and the number of days an employee is absent from work due to illness (Delgado & Kneisner, 1997).

Count data, in particular, consist of the observations or occurrences of behaviors (or similar dependent variables) observed within a fixed period of time that are positively-bounded between zero and infinity (Coxe, West, & Aiken, 2009; Stroup, 2013). Count outcomes are frequently considered positively skewed, because excessive zeros and additional low values can be present within the data set (Heck & Thomas, 2015). The presence of count data requires that alternative distributions be utilized (e.g., Poisson, Negative Binomial, etc.) to account for the true non-normality of the distribution.

Distributional Assumptions for Counts

Multilevel modeling assumes that data are normally distributed, which can be problematic when SCD outcomes are count and therefore, have a non-normal distribution. As previously mentioned, SCD data are rarely assumed to be normally distributed with the presence

of count data (Shadish et al., 2013). When normal distribution assumptions cannot be met, a different and more appropriate distributional assumption must be modeled. There are many distributions that a researcher can select based on theory, but the two most common distributions particularly for count data are the Poisson distribution, and the Negative Binomial distribution.

The introduction of non-normal distributions shifts the discussion from the traditional general linear model framework into the generalized linear model framework (GLiM). The GLiM framework provides a way to establish accurate results from binary, ordered categorical, and *count* data sets (Olsson, 2002). The GLiM framework modifies two components of the ordinary least squares (OLS) framework, such that (1) a nonlinear relationship between the dependent variable and the predictors can become linear by transforming the predicted outcome, and (2) the error structure is more flexible as compared to traditional OLS regression (Coxe et al., 2009).

The Poisson distribution originated from Poisson regression, which is subsumed under the umbrella of the generalized linear model framework (Coxe et al., 2009). Count data typically follow a Poisson distribution that involves only one parameter (λ) to represent both the mean and the variance. The mean and variance either increase or decrease together in a Poisson distribution, and therefore, only one parameter needs to be identified in MLM identification (Coxe, et al., 2009). The Poisson distribution uses a link function ($\eta = \ln(\lambda)$), which is a natural log link to transform the nonlinear relationship to a linear one (Heck & Thomas, 2015; Coxe et al., 2009).

When the mean and variance are equal, equidispersion is maintained. Conditions where equidispersion is not present, however, pose additional challenges to researchers (to be discussed). Compared to a normal distribution, the Poisson distribution is more appropriate for

handling the properties of count data, because this distribution can model the count outcomes due their integer values of zero or greater (zero and positively-bounded integers only). The Poisson regression model can, therefore, be represented in Equation 12 as

$$\ln(\hat{\mu}) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p, \qquad (12)$$

which contains the predicted count on the outcome variable ($\hat{\mu}$) conditional on the specific predictor values (X_1, X_2, so on), the intercept (b_0) and the regression coefficients (b_1). The major distinction between the traditional OLS regression and the Poisson regression is that the predicted score is the natural logarithm (ln) of the count, rather than the count itself (Coxe et al., 2009).

There are four alternative situations that may arise when equidispersion is not maintained; these situations are common in practice and fall within two major distinctions, dispersion-related and zero-related (Coxe et al., 2009). The dispersion-related situations refers to instances in which greater variability is present than would be expected within the count data, otherwise known as overdispersion. Conversely, underdispersion occurs when less variability is present than would be expected. The zero-related distinction refers to instances where zero values are either too few (truncated) or too many (excessive). These four Poisson-related situations (under- and overdispersion, truncated zeros, and excessive zeros) require the use of additional distribution specifications that correct for inaccurate estimates and invalid inferences.

The negative binomial distribution is an extension of the Poisson distribution in that the negative binomial accounts and corrects for overdispersion related to count data (Shadish et al., 2013). Failing to account for this overdispersion results in four main issues: (1) standard error estimates will be too small, (2) test statistics will be too large, (3) statistical significance will be

overestimated, and (4) confidence limits will be too small (Coxe et al., 2009). This, therefore, lends support to the argument that overdispersion is the biggest modeling issue for count data.

The negative binomial distribution can further correct for overdispersion by accounting for the heterogeneity and unexplained variability between individuals on the same predicted value, where individuals are still presented by a Poisson distribution with a different mean parameter (Coxe et al., 2009). The negative binominal distribution contains two parameters to represent the mean (m) and the dispersion parameter (k), such that the dispersion parameter measures the dispersion of the distribution (White & Bennetts, 1996). While the Poisson distribution assumes equidispersion, the negative binomial distribution does not require this assumption. As the variance of a negative binominal approaches the mean (m), or the overdispersion decreases, the dispersion parameter (k) approaches infinity and p approaches 0 (Bliss & Fisher, 1953), where p is m/k.

Present Study

The use of complex statistical methods, specifically multilevel modeling, for single-case designs has become increasingly prevalent in the literature. Previous research evaluating the analysis of count data and SCDs across various study conditions (e.g., time-series length, degree of freedom methods) have aided in advancing the utility of such complex statistical techniques (i.e., multilevel modeling). However, this previous research has failed to account for the non-normal nature of common recurrent event (count) outcomes (Shadish et al., 2013). Rather than assume count data are normally distributed, the present research considers the theoretically plausible distributions (i.e., Poisson distribution, negative binomial) that the data subsume.

The purpose of the present study is to evaluate the utility of multilevel modeling for handling recurrent event outcomes in synthesized single-case designs using Monte Carlo simulation across various analysis and data generation conditions. The following research questions were posed:

1. What effects do distributional assumptions (i.e., normal, Poisson, and negative binomial) have on relative bias, mean square error, and coverage of the mean estimators at each phase (i.e., baseline versus intervention)?
2. What effects do degree of freedom methods (i.e., between within, containment, residual, Kenward-Roger, Satterthwaite) have on relative bias, mean square error, and coverage for fixed effect phase mean estimators?
3. What effects do time-series length have on relative bias, mean square error, and coverage of mean estimators?

4. What effects do sample size at level 2 (e.g., student) and level 3 (e.g., teacher) have on relative bias, mean square error, and coverage of mean estimators? (Additional examples of level-2 and level-3 units on page 23.)
5. What, if any, effects do the interactions between the above conditions have on relative bias, mean square error, and coverage of mean estimators?

Based on previous research findings, the following hypotheses were formulated:

1. Adjusted distributional assumptions that are more appropriate for recurrent event (count) data (i.e., Poisson, negative binomial) will yield better simulation outcomes than ignoring the non-normality of counts (e.g., Coxe et al., 2009; Moghimbeigi, Eshraghian, Mohammad, & McArdle, 2008).
2. The Kenward-Roger degree of freedom method will outperform the other four degree of freedom methods under consideration, given its adjustment for small sample sizes (Ferron et al., 2009).
3. The length of the time series will not have a significant impact on the simulation outcomes (Ferron et al., 2009), absent of consideration for autocorrelation.
4. The “larger” sample sizes will yield better simulation outcomes (i.e., relative bias, mean square error, coverage) than the smaller sample sizes, given that multilevel modeling performs better with larger sample sizes.

Method

This two-stage research used (1) secondary data obtained through prior behavioral consultation research, and (2) simulated data through Monte Carlo simulation methods, such that, obtained parameter estimates in the first stage serve as population values in the second stage. The empirical context guided the decisions in both the multilevel modeling in stage one and the simulation study in stage two.

Stage 1: Multilevel Models with Empirical Context

Empirical Context

Conjoint Behavioral Consultation (CBC) is an indirect intervention that enables consultants to work collaboratively with parents and teachers to target students with academic, social, or behavioral needs (Sheridan, 1990; Sheridan & Kratochwill, 2008). The manifestation of behavioral problems across multiple settings (i.e., home and school) highlights the need for an intervention plan that connects aid and support from all individuals in those settings. In real-world applications, the ability to analyze recurrent event (count) outcomes (e.g., frequency of disruptive behaviors in the classroom) appropriately ensures that behavioral and clinical data are appropriately interpreted and proper treatment implementations are employed.

Participants

The secondary data that provided the empirical context for this study were collected from two large field-based randomized control trials (RCTs) with repeated measures design: CBC Early Grades (Sheridan, Bovaird, Glover, Garbacz, Witte, & Kwon, 2012) and CBC Rural Communities (Sheridan, Holmes, Coutts, & Smith, 2012; Sheridan, Holmes, Coutts, Smith, Kunz, & Witte, 2013). The CBC Early Grades data set was comprised of 157 elementary students and 74 teachers in a moderately-sized Midwestern city and surrounding communities.

The CBC Rural Communities data set was comprised of 224 elementary students and 133 teachers. There was missing data for 18 students which were removed from the final data set. The final sample size, therefore, consisted of 381 elementary students and 207 teachers based on complete data. In order to simplify the model, school-level information was not considered in the present study.

As part of the intervention and business-as-usual control conditions, each of the 381 elementary students participated in a small SCD consisting of repeated observations within the same child, measured over 10 measurement occasions. Since students began participation at different times (i.e., staggered start times), the collective sample of 381 students can effectively be considered a large non-concurrent, or natural, multiple baseline design (Harvey, May, & Kennedy, 2004). In the context of the broader CBC RCTs, synthesizing multiple individual-level SCDs allows the research team to consider inter-individual effects and potentially population-level inference. All intervention participants are measured on multiple occasions within a baseline phase and an experimental phase (See *Instrument and Procedures*), as well as serving as their own control in the baseline phase prior to the administration of the CBC intervention. Participants in the control condition provide additional true counterfactual information by not receiving the CBC intervention but are still measured on the same schedule. Therefore, the multiple SCDs within the context of the CBC RCTs meet the necessary components of single-case designs, as outlined by Kazdin (2011): SCDs have continuous assessment over time, and the same subject is used as their own control over time (p. 384 – 385).

Instrument and Procedures

The RCTs were comprised of a control condition and an experimental condition. The control condition was "business-as-usual," whereas the experimental condition consisted of a

CBC intervention within a group setting of one to three children per one teacher. The experimental and control conditions utilized the Parent Daily Report (PDR; Chamberlain & Reid, 1987), which consisted of a 34-item daily observation and self-report measure. Using the PDR, parents recorded the frequency of disruptive behaviors that occurred in the last 24 hours. Parents recorded these observations from a list consisting of all 34 internalizing and externalizing disruptive behaviors. See *Table A1* for a complete list of the disruptive behaviors. The PDR was completed ten times over a 5-10 week period, which included four observations for the control condition and six observations for the experimental condition. At each of the 10 observation points, parents recorded the presence or absence of the 34 disruptive behaviors. Chamberlain and Reid (1987) reported an inter-interviewer reliability, $r = 0.98$.

Data Synthesization

The current study merged together data from the two RCTs (CBC Rural and CBC Early Grades). Certain information such as participant identification and condition assignment (control versus CBC) were integrated from separate databases into the final data set for this study. The 34 disruptive behaviors of the PDR were coded as either the behavior did occur (code = 1) or the behavior did not occur (code = 0). A count variable was included to sum the total number of disruptive behaviors at each of the 10 observation points for the PDR, such that the count variable ranged from 0 to 34. The frequency (or count) of these disruptive behaviors was the dependent variable for this study.

Multilevel Models

Theoretical Model

A single multilevel model was estimated based on the (a) proposed theoretical model and the (b) data-driven model based on the "best-fitting model" derived from the secondary data.

The theoretical model posited that a 3-level MLM would be most appropriate, due to the presence of observations nested within students and students nested within teachers. The theoretical model was proposed prior to data analysis and was also the basis for the forthcoming simulation study.

The three-level MLM was fit to the data, varying on the type of recurrent event distributional assumptions, with repeated observations (level 1), students (level 2), and teachers (level 3). There were three distributional assumptions: (a) normal distribution, (b) Poisson distribution, and (c) Negative Binomial distribution. Five degree of freedom methods were also varied across the distribution assumptions, which included (a) containment, (b) residual, (c) between-within, (d) Satterthwaite, (e) Kenward-Roger.

The level-1 regression equation (Equation 13) for repeated observations nested within students was modeled as:

$$\eta_{ijk} = \pi_{0jk} + \pi_{1jk} Phase_{ijk} \qquad (13)$$

The occasions (or measurement occasions) nested within students was modeled for the second level in Equations 14 and 15 as:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \qquad (14)$$

and

$$\pi_{1jk} = \beta_{10k} + r_{1jk} \qquad (15)$$

At the third level, students nested within teachers was modeled in Equations 16 and 17 as:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad (16)$$

and

$$\beta_{10k} = \gamma_{100} + u_{10k} \qquad (17)$$

The combined model was represented as: $\eta_{ijk} = \gamma_{000} + \gamma_{100}Phase_{ijk} + r_{0jk} + r_{1jk}Phase_{ijk} + u_{00k} + u_{10k}Phase_{ijk}$, where γ_{000} is the intercept, $\gamma_{100}Phase_{ijk}$ is the phase effect, r_{0jk} is the level-2 random intercept, r_{1jk} is the level-2 random slope, u_{00k} is the level-3 random intercept, and u_{10k} is the level-3 random slope.

$$\eta_{ijk} = log(\mu_{ijk}) \qquad (18)$$

$$\mu_{ijk} = E(Y_{ijk}) \qquad (19)$$

where Y_{ijk} is the number of disruptive behaviors observed at occasion (*i*) for child (*j*) in teacher (*k*). It is also assumed that:

$$Y_{ijk} \mid u \sim Poisson(\mu_{ijk} \mid u) \quad (20)$$

Where $\begin{bmatrix} r_{0jk} \\ r_{1jk} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\pi 00} & 0 \\ 0 & \tau_{\pi 10} \end{bmatrix}\right)$ & $\begin{bmatrix} u_{00k} \\ u_{10k} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\beta 00} & 0 \\ 0 & \tau_{\beta 10} \end{bmatrix}\right)$; it is further assumed that the random effects are uncorrelated across levels.

The theoretical MLM was estimated using the Restricted Maximum Likelihood (REML) estimator via PROC GLIMMIX in SAS version 9.4 (SAS Institute Inc., 2013) for models that assumed normal distribution. The models that assumed the Poisson and Negative Binomial distributions were estimated with the Residual Pseudo-likelihood (RSPL) estimator. To avoid convergence issues, the Newton-Raphson with Ridging optimization technique was implemented for the Negative Binomial distribution, rather than the default option the Dual Quasi-Newton (Kiernan, Tao, & Gibbs, 2012). The Newton-Raphson technique is considered most appropriate when dealing with overdispersion in the negative binomial distribution.

Data-Driven Model

The data-driven model was also estimated to determine the "best fitting" model to be derived from the secondary CBC data. The data-driven model was estimated using the

Maximum Likelihood (ML) estimator with the Laplace likelihood approximation. The Laplace likelihood approximation allowed for comparisons of model fit that would have otherwise been unavailable in pseudo-likelihood estimators. The data-driven model varied across distributional assumption on the count outcome (normal, Poisson, and Negative Binomial). Similar to the theoretical model, the Newton-Raphson with Ridging optimization was implemented. The results from the data-driven model were not used in the stage two simulation.

Justification

The decision to accept and continue with the theoretical model was based on the theory-based design, in which repeated observations (level 1) are nested within student (level 2) nested within teacher (level 3). The sole purpose of the theoretical model was to inform the forthcoming simulation based on previous research (Bovaird, Sheridan, & Glover, 2009). The addition of the control group (no CBC intervention) for the final data set was included in favor of (a) increasing the level-3 units and (b) simplifying the model; this resulted in the inability to capture treatment effects within this study. Therefore, the findings for this study should not be used for empirical purposes (to be discussed in limitations).

Stage 2: Monte Carlo Simulation

Design

Monte Carlo simulation studies "allow researchers to assess the finite sampling performance of estimators by creating controlled conditions from which sampling distributions of parameter estimates are produced" (Paxton, Curran, Bollen, Kirby, & Chen, 2001). The current study utilized Monte Carlo simulation methods to investigate the performance of the fixed effect phase mean estimators.

This study utilized a 5 x 4 x 3 x 2 factorial design for data generation and analysis conditions. The factorial design included four independent variables: (a) number of participants at level-2 (12, 30, 60, and 90) and number of participants at level-3 (6, 15, 30, and 45) with a constant ratio of students to teachers, (b) length of the time series (10, 20, and 30), and (c) distributional assumptions in which data were generated based on two distributions (Poisson and Negative Binomial). More specifically, Poisson-generated data were analyzed with Poisson and normal distributions, and Negative Binomial-generated data were analyzed with Negative Binomial distribution. One thousand replications were simulated for each of the 120 conditions using R (R Core Team, 2015). The parameter estimates obtained from Stage 1 (see Tables B1 – B3) were used as population values for the simulation study; such that, simulated data were generated through R (R Core Team, 2015). The dependent variables were relative bias, mean square error, and coverage of the fixed effect phase mean estimators.

Conditions Sampled

Number of participants. The number of participants at level-2 had four levels (12, 30, 60, and 90). The number of participants at level-3 also had four levels (6, 15, 30, and 45). The cluster size for students was held constant at two; the ratio was fixed at two students per one teacher across conditions based on the theoretical and empirical design. These levels were chosen based on previous empirical research. For example, Van den Noortgate and Onghena (2008) presented an empirical example that included 30 participants across studies for a multilevel meta-analysis of raw data from SCDs. Owens and Ferron (2012) simulated data for four and eight participants to represent the high and low bounds for level-2 sample sizes. The empirical context for this study dictated the inclusion of $n = 60$ for consistency purposes within

the simulation study. The remaining levels of sample size were retained to assess study outcomes based on increases of participants at each level.

Time Series Length. The length of time series consisted of three levels: 10, 20, and 30 observations with a baseline phase (40% of observations) and a treatment phase (60% of observations). The 40/60 ratio of baseline and treatment phases design was based on the CBC empirical context, in which the time-series length of 10 included 4 baseline observations and 6 treatment observations. The time-series length of 20 included 8 baseline and 12 treatment observations. The time-series length of 30 consisted of 12 baseline and 18 treatment observations. Previous research recommends a minimum of five baseline observations to establish trends (The WWC Standards; Kratchowill et al., 2010), and a median of 20 data points based on the meta-analysis of 809 single-case designs implemented within 113 studies (Shadish & Sullivan, 2011).

Theoretical Distributions. The theoretical distributions were simulated at two levels: Poisson and Negative Binomial. Single-case design research commonly utilizes normal or normal with sandwich estimator distributions, despite the prevalence of recurrent event (count) outcomes. The Poisson distribution was included as a simulation condition to account for distributions in which the mean is equal to the variance. The Negative Binomial distribution was included to account for overdispersion, or instances in which the variance of the outcome was larger than the mean of the outcome. These two instances are predominant within recurrent event (count) outcomes. The normal distribution assumption was included in the analysis to demonstrate the inappropriateness of normal distribution when count outcomes are observed (Shadsih & Sullivan, 2011).

Data Generation and Analysis

Data were generated in R (R Core Team, 2015) based on the three-level theoretical model posited in Stage 1 (see Equations 1 - 5). The population parameter estimates obtained in Stage 1 (see Tables B1 – B3) were used to generate the data for the Poisson distribution and Negative Binomial distribution. The generated conditions were measurement observations (*M = 10, 20, and 30*), student sample size (*S = 12, 30, 60, and 90*) and teacher sample size (*C = 6, 15, 30, and 45*), resulting in 24 conditions across both distributions. The level-3 intercept (β_{00}) and level-3 phase effect (β_{10}) had negative variances, which required that the unconditional intraclass (ICC) correlations be calculated in order to correct for the negative values (Hedges & Hedberg, 2007).

One thousand replications per condition were generated for Poisson and Negative Binomial distributions. The generated data were then analyzed in SAS version 9.4 (SAS Institute Inc., 2013) based on the same theoretical model as in Stage 1. The outcome criteria for evaluating the model performance were mean square error, relative bias, and coverage. Population parameters were scaled from the log scale to the data scale to provide reliable, consistent estimates and interpretation in the normal distribution condition (Stroop, 2012).

Results

This section presents the results of the two-stage research design with the multilevel model results from stage one and the current simulation study from stage two. The theoretical model and data-driven model results are presented first to show model fit and corresponding parameter estimates to compare the two models; the theoretical model provided the basis for the simulation study in stage two.

Stage 1: Theoretical Multilevel Model

A three-level multilevel model was conducted on $n = 381$ elementary students within $k = 207$ teachers to determine the parameter estimates (*Tables B1 and B2*) based on the theoretical model for the simulation study (stage two). The parameter estimates in the theoretical model across the Poisson and Negative Binomial distributions were used to generate data in the simulation study. The level-1 variable included the count variable (counts) as the dependent variable. The maximum number of iterations had to be increased from the default of 20 iterations to 100 iterations (normal distribution) and 1,000 iterations (Poisson and Negative Binomial distributions) in order to facilitate successful convergence across distributions.

The distributions were varied across normal, normal with sandwich estimator, Poisson, and Negative Binomial. All five denominator degree of freedom methods (i.e., between-within, containment, residual, Kenward-Roger, and Satterthwaite) were also varied across distributional assumptions. Kenward-Roger and Satterthwaite were unavailable for the normal distribution with sandwich estimators, but the other three methods (between-within, residual, and containment) were implemented. The estimation technique for the normal distribution and normal distribution with sandwich estimators was set to restricted maximum likelihood, whereas the estimation technique for the Poisson distribution and Negative Binomial distribution was set

to restricted pseudo-likelihood. These estimation techniques were kept consistent across stage one and stage two.

Assessing AIC and BIC across the distributional assumptions (*Table B4*) indicated that the Negative Binomial distribution was the best-fitting model, based on the lowest AIC and BIC estimates. The normal distribution and normal distribution with sandwich estimator had substantially higher AIC and BIC than Poisson and Negative Binomial distributions.

Stage 1: Data-Driven Multilevel Model

The purpose of the data-driven multilevel model was to determine the "best fitting model" based on the empirical data set, which served as the empirical basis for this study. The process of determining the best-fitting model involved fitting two primary models to the data, which began by estimating an empty model without random effects (Model 1a). The next model consisted of estimating the empty model with a level-2 random intercept (Model 1b). The log likelihood difference between Model 1a and Model 1b indicated a significant difference, $p < .0001$, suggesting that Model 1b fits the empirical data better than Model 1a. The next model tested was the empty model with a level-3 random intercept (Model 2), which was compared to Model 1b. The log likelihood difference test between Model 1b and Model 2 indicated a non-significant difference, $p > 0.05$, suggesting that Model 2 does not fit the empirical data set better than Model 1b. See *Tables C1 – C4* in Appendix C for the model results.

The model fit was assessed using AIC and BIC across the two-level data-driven model and the three-level theoretical model (*see Table 1*). Results indicated that the three-level theoretical model produced lower AIC and BIC estimates for the Negative Binomial distribution, which were lower than the two-level data-driven model under the Negative Binomial distribution. The estimation technique for the data-driven model was set to Maximum Likelihood with the Laplace likelihood approximation to receive model fit information (i.e., AIC

and BIC). Model fit was assessed across normal, Poisson, and Negative Binomial distributions for the data-driven model. The Negative Binomial distribution resulted in lower values for AIC and BIC, compared to the other two distributions.

Justification – Part Two

The purpose of the two-level data-driven model was to determine the best fitting model based on the actual empirical data set. In other words, the data-driven model was compared to the theoretical model in order to assess which model was "best" for the simulation study in stage two. The data-driven model and the theoretical model clearly deviate from one another, so the decision to continue with the theoretical model was based solely on the theory behind the data (i.e., there are three levels to account for). More specifically, the theory was not disregarded simply because the empirical data set was not consistent with the three-level theoretical model. The theoretical model also produced AIC and BIC values that were lower than the data-driven model (see *Table 1*).

Stage 1: Results and Interpretation

It is important to stress that the following results from the two models are provided solely for illustration purposes. The results from the analyses should not be interpreted in any real manner due to the incorporation of both the control group and the treatment group, thus preventing any real treatment effect. The goal of the theoretical model was simply to provide population parameter estimates for the forthcoming simulation, whereas the purpose of the data-driven model was simply to acknowledge that the current empirical data do not agree with the theoretical model.

The results from the theoretical model indicate that the Negative Binomial distribution has the "best fit" in terms of AIC and BIC (see *Table B4*). The results indicate that the average

number of disruptive behaviors (γ_{000} = 1.917) at the beginning of the study is significant, $t_{(164.9)}$ = 56.08, p < .0001. There was also a significant decrease (γ_{100} = -0.382) in the number of disruptive behaviors for the shift in phase from control to treatment, $t_{(175.6)}$ = -15.27, p < .0001. The student-level random intercept (r_{0jk} = 0.4031) and the student-level random slope (r_{1jk} = 0.1558) were also significant, p < .05, indicating that there is significant variability between students at the beginning of the study and there is significant variability between students across phases, respectively. The teacher-level random intercept (u_{00k} = -0.0153) and teacher-level random slope (u_{10k} = -0.026) indicate that there is nonsignificant (p > .05) variability between teachers at the beginning of the study or across phases, respectively.

The results from the theoretical model differed when the normal distribution was implemented, such that the fixed effects (γ_{000} = 7.998, γ_{100} = -2.108) were still statistically significant, p < .0001, but the student-level intercept and slope (r_{0jk} = 17.855 and r_{1jk} = 3.507) and teacher-level intercept and slope (u_{00k} = -1.563 and u_{10k} = -0.487) were nonsignificant, p > .05, indicating that there is no variability between students or teachers at the beginning of the study or across phases. See *Tables B1 – B2* for theoretical model parameters.

The results from the data-driven model indicate that the Negative Binomial distribution has the "best fit" in terms of AIC and BIC (see *Table 1*). The results indicate that the average number of disruptive behaviors (γ_{000} = 1.906) at the beginning of the study is significant, $t_{(380)}$ = 50.78, p < .0001. There was also a significant decrease (γ_{100} = -0.347) in the number of disruptive behaviors for the shift in phase from control to treatment, $t_{(3071)}$ = -17.94, p < .0001. The student-level random intercept (r_{0jk} = 0.452) was also significant, p < .05, indicating that there is significant variability between students at the beginning of the study, which is consistent with the interpretation of the theoretical model. The data-driven model produced a similar

interpretation for the student-level intercept and slope within the normal distribution as in the theoretical model. See *Tables C1 – C3* for data-driven model parameters.

Assessing the model fit between the theoretical model and data-driven model, the theoretical model was selected for the simulation in stage two (see *Table 1*) based on the theoretical model producing the lowest AIC and BIC values and alignment with the theory behind the research. The model results for the theoretical model were then used as the population parameter estimates to inform the simulation work in stage two (see *Table B3*).

Table 1.

Comparison of Fit Statistics between Distributional Assumptions

		AIC	AICC	BIC
Data-Driven	Normal	19307.10	19307.11	19322.87
	Poisson	19021.74	19021.75	19033.57
	Negative Binomial	18406.37	18406.38	18422.14
Theoretical	Normal	19259.29	19259.29	19271.12
	Normal w. Sandwich	19261.11	19261.13	19277.77
	Poisson	18653.54	18653.54	18669.31
	Negative Binomial	18299.17	18309.17	18309.19

Note: The 2-level data-driven model and the 3-level theoretical model, respectively.

Stage 2: Monte Carlo Simulation and Analysis

The second stage of the research presents the results from the simulation and analysis based on the theoretical model. The initial results are organized based on research questions presented. A more meaningful discussion is presented in the last research question, which focused on the interaction between the conditions (distributional assumption, degree of freedom methods, time-series length, and sample size). The subsequent discussion also refers to "baseline

phase" (i.e., "Phase 0") to refer to the control condition, and "intervention phase" (i.e., "Phase 1") to refer to the experimental condition (as outlined in the *Instrument and Procedures* section).

Research Question #1:

What effect does distributional assumption have on relative bias, mean square error, and coverage of the mean estimators at each phase?

The extent to which the fixed effects were biased was assessed based on the acceptable absolute levels of relative bias (less than 0.05; Hoogland & Boomsma, 1998). The results of the normal distribution indicated that, independent of other simulation conditions, the relative bias levels exceeded 0.05 for the baseline phase (*min = 0.262, max = 0.268*) with increased bias in the intervention phase (*min = 0.399, max = 0.419*). The Poisson distribution never exceeded the acceptable level of relative bias for the baseline phase (*min = 0.019, max = 0.034*) with minor increases in bias for the intervention phase (*min = 0.029, max = 0.039*). Similarly, the Negative Binomial distribution also never exceeded the acceptable level of relative bias for the baseline phase (*min = -0.003, max = - 0.006*) with minor increases in the bias for the intervention phase (*min = -0.004, max = 0.0109*). Across the three distributions, the results indicated that the Negative Binomial distribution displayed the lowest levels of relative bias at each phase.

The extent to which the mean square error for the fixed effects performed was assessed based on the relative levels of mean square error, such that estimators with lower levels of error are preferred over estimators with higher levels of error. The lower levels produce estimates that are closer to the population parameter (Koziol, 2015). The results indicated that, independent of other study conditions, the mean square error was substantially higher in the normal distribution compared to the Poisson and Negative Binomial distributions. The normal distribution displayed substantially higher levels of mean square error for the baseline phase (*min = 3.363, max = 4.64*)

with increases in the intervention phase (*min = 3.67, max = 5.09*). The Poisson distribution resulted in overall lower levels of mean square error for the baseline phase (*min = 0.102, max = 0.833*) with decreases in the intervention phase (*min = 0.071, max = 0.547*). The Negative Binomial distribution performed better than the other two distributions in terms of mean square error with lower levels in the baseline phase (*min = 0.004, max = 0.016*) with minor increases in the intervention phase (*min = 0.004, max = 0.024*).

The extent to which the confidence interval coverage for the fixed effects performed was assessed based on the nominal coverage rate of 95%, such that estimators with greater than .95 coverage rates are considered too conservative (overcoverage) and estimators with less than .95 coverage rates are considered too liberal (undercoverage; Agresti & Caffo, 2000). The confidence interval coverage rates for the normal distribution tended to undercover for the baseline phase (*min = 0.001, max = .791*), with increases in the intervention phase (*min = 0.001, max = 0.851*). The confidence interval coverage rates for the Poisson distribution tended to undercover and overcover in the baseline phase (*min = 0.899, max = 0.964*) with increases in the intervention phase (*min = 0.903, max = 0.972*). The undercoverage and overcoverage vary based on other study conditions (See Research Question #5). The confidence interval coverage rates for the Negative Binomial distribution tended to undercover and overcover in the baseline phase (*min = .903, max = .959*) and in the intervention phase (*min = .907, max = .952*). The Negative Binomial distribution had fewer instances of overcoverage than the Poisson distribution.

Summary: The adjusted distributional assumptions, Poisson and Negative Binomial, tended to have better levels of relative bias and mean square error across phases and tended to

have confidence interval coverage rates closer to the nominal .95 acceptable level, with Negative Binomial performing better than Poisson overall.

Research Question #2:

What effect does degree of freedom method have on relative bias, mean square error, and coverage for fixed effect phase mean estimators?

The results of the degree of freedom methods indicated that varying the method had no impact on relative bias levels in the baseline phase (*min = -0.006, max = 0.356*) with increases in the intervention phase (*min = -0.0063, max = 0.419*), independent of other simulation conditions. The interaction between degree of freedom methods and distributional assumption, time-series length, and sample size has a greater impact on relative bias than the degree of freedom method alone.

Similar to relative bias, the degree of freedom methods had no impact on mean square error levels in the baseline phase (*min = 0.0047, max = 4.64*) with increases in the intervention phase (*min = 0.0068, max = 5.08*), independent of other simulation conditions. The degree of freedom methods had more of an impact when the remaining simulation conditions were also considered. See Research Question #5 for a discussion of the interactions between the degree of freedom methods and the remaining simulation conditions.

The confidence interval coverage rates tended to undercover for the between-within and residual degree of freedom methods in the baseline phase (*min = 0.898, max = 0.937*) and for the intervention phase (*min = 0.882, max = 0.935*), independent of other conditions. The between-within and residual methods were considered comparable to one another. The coverage rates tended to undercover drastically for the containment method, which consistently resulted in a coverage rate of zero. The coverage rates tended to undercover and overcover for the Kenward-

Roger and Satterthwaite methods in the baseline phase (*min = 0.925, max = 0.955*) and in the intervention phase (*min = 0.906, max = 0.955*) with coverage rates closer to the nominal level of .95. The Kenward-Roger and Satterthwaite methods yielded almost identical coverage rates, due to the Kenward-Roger being derived from the Satterthwaite to account for unbalanced cluster sizes. The current research utilized balanced cluster sizes, so the differences in coverage rates between the two methods were further minimized. The confidence interval coverage rates for this research question are specifically for Poisson and Negative Binomial distributions, due to the substantially lower coverage rates that resulted in the normal distribution (See *Figure V3*).

Summary: The confidence interval coverage rates varied considerably across the five degree of freedom methods. The Kenward-Roger and Satterthwaite methods produced coverage rates that were closer to the nominal level of .95.

Research Question #3:

What effect does time-series length have on relative bias, mean square error, and coverage of the mean estimators?

The results of the shortest time-series length of 10 observations indicated that, independent of other conditions, exceeded the 0.05 level of relative bias in some conditions and did not exceed the level in other conditions in the baseline phase (*min = -0.006, max = 0.27*) with increases in the intervention phase (*min = -0.006, max = 0.414*). The second time-series length of 20 observations resulted in a similar pattern where relative bias levels exceeded and did not exceed the 0.05 level under certain conditions in the baseline phase (*min = -0.006, max = 0.266*) and for the intervention phase (*min = -0.006, max = 0.419*). The longest time-series length of 30 observations also resulted in the same process in the baseline phase (*min = -0.006, max = 0.267*) with increases in the intervention phase (*min = -0.006, max = 0.413*). Overall, the

relative bias levels for the three time-series lengths were comparable to one another, with the maximum bias levels across phases in the normal distribution. The relative bias levels were, on average, under the acceptable 0.05 level for the Poisson and Negative Binomial distributions, where the normal distribution consistently exceeded the acceptable relative bias level. See Research Question #5 for more discussion on these interactions.

The results of the shortest time-series length of 10 observations and the second time-series length of 20 observations indicated that there were relatively similar mean square error levels in the baseline phase (*min = 0.002, max = 4.64*) with increases in the intervention phase (*min = 0.006, max = 5.08*). The results of the longest time-series length of 30 observations indicated that there were smaller maximum mean square error levels in the baseline phase (*min = 0.004, max = 4.42*) and in the intervention phase (*min = 0.007, max = 4.42*). Overall, the time-series lengths were comparable to one another on the mean square error levels.

The confidence interval coverage rates for the shortest time-series of 10 observations generally tended to undercover in the baseline phase (*min = .898, max = .951*) and for the intervention phase (*min = .882, max = .955*), with maximum estimates close to the nominal .95 acceptable level. The confidence interval coverage rates for the second time-series length of 20 observations generally tended to undercover in baseline phase (*min = .908, max = .959*) and for the intervention phase (*min = .907, max = .952*), with maximum estimates close to the acceptable level. The confidence interval coverage rates for the longest time-series length of 30 observations tended to undercover and overcover in baseline phase (*min = .903, max = .964*) with increases in the intervention phase (*min = .914, max = .971*). The overcoverage for the longest time-series length was much more substantially than in the other two time-series lengths. The confidence interval coverage rates were based on the Poisson and Negative Binomial

distributions, due to the normal distribution resulting in considerable undercoverage in most cases (*min = .001, max = .788*). Overall, the confidence interval coverage rates increased, on average, with the increase in time-series length.

Summary: The manipulation of time-series lengths minimally affected relative bias and mean square error levels. The second time-series length of 20 observations resulted in coverage rates that were, on average, closer to the nominal level of .95, with the longest time-series length of 30 observations resulting in overcoverage in most cases.

Research Question #4:

What effect does sample size at level 2 and level 3 have on relative bias, mean square error, and coverage of the mean estimators?

The results of varying the sample sizes indicated that the relative bias level for the smallest sample size (*j = 12, k = 6*) exceeded the 0.05 in some cases and not in other cases in the baseline phase (*min = -0.006, max = .266*) with increases in the intervention phase (*min = -0.006, max = .414*). The second sample size (*j = 15, k = 30*) similarly exceeded the 0.05 in some cases and did not exceed 0.05 in other cases in the baseline phase (*min = -0.006, max = .265*) with increases in the intervention phase (*min = -0.006, max = .414*). The largest sample size (*j = 60, k = 30*) also exceeded the 0.05 in some cases and not in others in the baseline phase (*min = -0.006, max = 0.267*) with increases in the intervention phase (*min = -0.006, max = 0.412*). The Poisson and Negative Binomial distributions resulted in relative bias that did not exceed the 0.05 acceptable level, whereas the normal distribution consistently exceeded 0.05. Overall, the three sample sizes at level-2 and level-3 were comparable to one another on relative bias levels.

The results indicated that the smallest sample size (*j = 12, k = 6*) had higher mean square error levels than the other two sample sizes, for the baseline phase (*min = 0.004, max = 4.64*) with increases in the intervention phase (*min = 0.007, max = 5.089*). The second sample size (*j = 30, k = 15*) had lower mean square error levels in the baseline phase (m*in = 0.0049, max = 3.84*) with increases in the intervention phase (*min = -0.006, max = 4.064*). The largest sample size (j = 60, k = 30) had even lower mean square error levels in the baseline phase (*min = 0.002, max = 3.5*) with increases in the intervention phase (*min = 0.003, max = 3.79*). The distributional assumption influenced the mean square error levels for each sample size; therefore, the interaction indicates more meaningful results (See Research Question #5). Overall, the mean square error levels varied based on the sample size at level-2 and level-3, with lower mean square error levels represented in the largest sample size.

The confidence interval coverage rates for the smallest sample size tended to undercover in the baseline phase (*min = .898, max = .903*) with increases in coverage in the intervention phase (*min = .906, max = .955*). The coverage rates for the second sample size tended to undercover in the baseline phase (*min = .898, max = .954*) and in the intervention phase (*min = .906, max = .952*). The confidence interval coverage rates for the largest sample also tended to undercover in the baseline phase (*min = .912, max = .952*) and in the intervention phase (min = *.882, max = .956*). The coverage rates presented in this section were based on the Poisson and Negative Binomial distributions; the normal distribution tended to substantially undercover (e.g., min = 0.001, max = 0.791) across sample sizes.

Summary: The sample sizes at level-2 and level-3 had little impact on relative bias levels, whereas the largest sample size resulted in lower mean square error levels and the smallest sample size resulted in the highest mean square error levels. The confidence interval coverage

rates tended to be undercovered across sample sizes with minor increases as sample size increases.

Research Question #5:

What, if any, effects do the interactions between the above conditions have on relative bias, mean square error, and coverage of the mean estimators?

The interaction(s) between the four conditions (degree of freedom methods, sample size, time-series length, and distributional assumptions) provided more meaningful findings than the individual conditions alone. Research studies often utilize multiple conditions, and therefore, a single condition may produce less information than the interaction between conditions.

Relative Bias across Theoretical Distributions, Dependent on Sample Sizes and Time-Series Lengths

Normal Distribution

The results of the normal distribution indicated that the inclusion of sample size and time-series length had minimal impact on the relative bias levels. More specifically, the normal distribution consistently resulted in relative bias levels that exceeded 0.05, with increases in relative bias from the baseline phase to the intervention phase (See *Figure 1*). The three sample size conditions resulted in almost identical relative bias levels across the three time-series lengths of 10, 20, and 30 observations in the baseline phase (*min = 0.263, max = 0.267*) with increases in the intervention phase (*min = 0.399, max = 0.419*). The time-series lengths also had minimal impact on the relative bias levels (e.g., *RB ≈ 0.263, 0.266, 0.266* for 10, 20, and 30 observations, respectively), and this pattern was consistent across sample sizes. The relative bias levels that consistently exceeded the acceptable 0.05 supports that the normal distribution does not perform well overall, and does not improve with the variations of sample size and time-series lengths.

Figure 1. Relative bias across sample size and time-series lengths for the normal distribution. There is an increase in relative bias levels from baseline to intervention phase across each conditions. The relative bias levels are comparable to each other across sample sizes and time-series lengths, and exceeded the acceptable level of 0.05. See *Table 2* for key to interpretation.

Table 2.

Key for Figures across Sample Sizes and Time-Series Length

Code (in figure)	**Level-two sample size**	**Level-three sample size**	**Length of time series**
6.12.10	*j = 12*	*k = 6*	10 observations
15.30.10	*j = 30*	*k = 15*	10 observations
30.60.10	*j = 60*	*k = 30*	10 observations
6.12.20	*j = 12*	*k = 6*	20 observations
15.30.20	*j = 30*	*k = 15*	20 observations
30.60.20	*j = 60*	*k = 30*	20 observations
6.12.30	*j = 12*	*k = 6*	30 observations
15.30.30	*j = 30*	*k = 15*	30 observations
30.60.30	*j = 60*	*k = 30*	30 observations

Note: for Figures 1 – 8.

Poisson Distribution

The Poisson distribution performed substantially better in regards to relative bias than the normal distribution. The results of the Poisson distribution indicated that the inclusion of sample size and time-series length had more of an influence on the levels of relative bias. The Poisson distribution never exceeded the acceptable level of 0.05 for relative bias across all sample sizes and time-series lengths (see *Figure 2*). More specifically, the second sample size (*j = 15, k = 30*) produced higher relative bias across time-series lengths, but still maintained levels less than 0.05 in the baseline phase (*min = 0.267, max = 0.034*) with increases in the intervention phase (*min = 0.03, max = 0.398*). The smallest sample size (*j = 12, k = 6*) and largest sample size (*j = 60, k = 30*) produced similar levels of relative bias for the time-series length of 10 observations in the baseline phase (*min = 0.026, max = 0.027*) with minor increase in the intervention phase (*min = 0.028, max = 0.032*).

The lowest level of relative bias was observed in the time-series length of 30 observations with the smallest sample size; such that, the relative bias level was the closest to zero in the baseline phase (0.017) and in the intervention phase (0.021). It should be noted that there is a pattern inconsistency for the time-series length of 20 observations from the other two observations, in that the smallest sample size increases more drastically. This inconsistency may be due to a convergence issue, but currently warrants additional investigations. Overall, time-series length had minimal impact on relative bias levels.

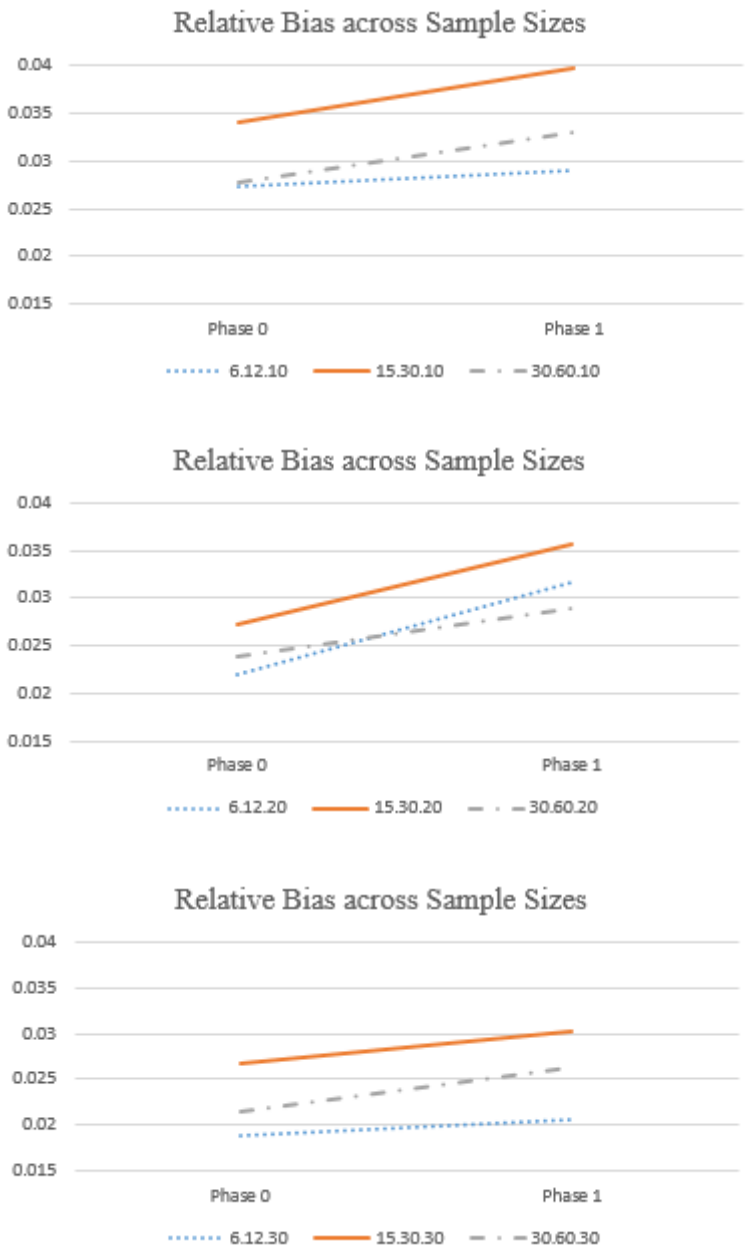

Figure 2. Relative bias across sample sizes and time-series lengths for the Poisson distribution. The second sample size ($j = 30$, $k = 15$) displayed higher levels of relative bias. The smallest sample size ($j = 12$, $k = 6$) performed better, on average, across time-series lengths. The relative bias estimates never exceeded the acceptable 0.05 level for relative bias across all sample sizes and time-series lengths. See *Table 2* for key to interpret figures.

Negative Binomial Distribution

The Negative Binomial distribution performed substantially better in regards to relative bias than the normal distribution, with even lower levels of relative bias than the Poisson distribution. The results of the Negative Binomial distribution indicated that the inclusion of sample size and time-series length influenced the levels of relative bias. The Negative Binomial distribution never exceeded 0.05 across all sample sizes and time-series lengths (see *Figure 3*). More specifically, the smallest sample size (*j = 12, k = 6*) produced the lowest levels of relative bias across the three time-series lengths for the baseline phase (*min = -0.006, max = -0.0059*) and in the intervention phase (*min = -0.006, max = -0.0103*). The largest sample size (*j = 60, k = 30*) produced the highest levels of relative bias across the three time-series lengths for the baseline phase (*min = -0.003, max = -0.0027*) and in the intervention phase (*min = -0.0037, max = -0.005*).

The shortest time-series length of 10 observations produced the highest levels of relative bias levels across the time-series lengths. The relative bias levels did not exceed the 0.05 acceptable level across all conditions. It should be noted that there is a pattern inconsistency in the time-series length of 10 observations, in which the largest sample size slightly increases as opposed to decreasing similar to the other time-series lengths. This inconsistency may be due to a misfit between the largest sample size and shortest time-series length or a convergence issue, but warrants more investigation. See *Figure 4* for a comparison of relative bias levels between normal distribution, Poisson distribution, and Negative Binomial distribution.

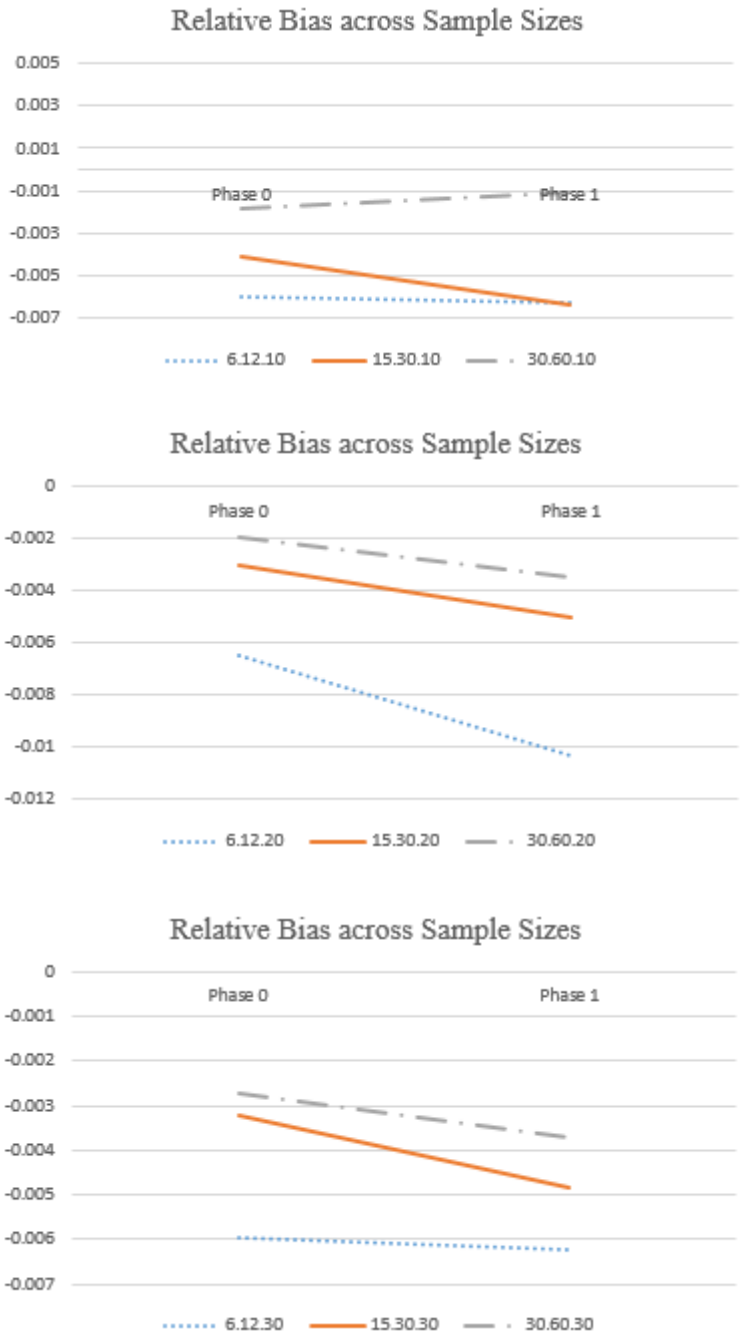

Figure 3. Relative bias across sample sizes and time-series lengths for the Negative Binomial distribution. The smallest sample size resulted in the lowest levels of relative bias across the three time-series lengths. The relative bias levels never exceeded 0.05 for the Negative Binomial distribution. Note: There is a pattern inconsistency of the largest sample size for the shortest time-series length. See *Table 2* for key to interpret figures.

Figure 4. Comparison of relative bias across distributional assumptions, sample sizes, and time-series lengths of 10, 20, and 30 observations, respectively. Note: the y-axes scaling is different in order to display relative bias levels clearly across distributions.

Mean Square Error across Theoretical Distribution, Dependent on Sample Sizes and Time-Series Length

Normal Distribution

The normal distribution had considerably high levels of mean square error across sample sizes and time-series lengths. The inclusion of sample size produced varying mean square error levels, with the largest sample size (*j = 60, k = 30*) producing the least amount of mean square error in the baseline phase (*min = 3.36, max = 3.51*) with increases in the intervention phase (*min = 3.67, max = 3.82*). The smallest sample size (*j = 12, k = 6*) produced the higher mean square error levels in the baseline phase (*min = 4.42, max = 4.64*) with increases in the intervention phase (*min = 4.42, max = 5.09*). The second sample size fell between the largest and smallest sample sizes across the three time-series lengths.

As sample size increased in the normal distribution, the mean square error levels decreased (see *Figure 5*). The time-series lengths produced relatively similar mean square error levels as the lengths increased from 10 observations to 30 observations, with the largest time-series length producing more stable mean square error levels across sample sizes (e.g., $MSE \approx 3.5 – 4.42$ versus $MSE \approx 3.36 – 5.09$ and $MSE \approx 3.67 – 5.09$ for 10 and 20 observations, respectively).

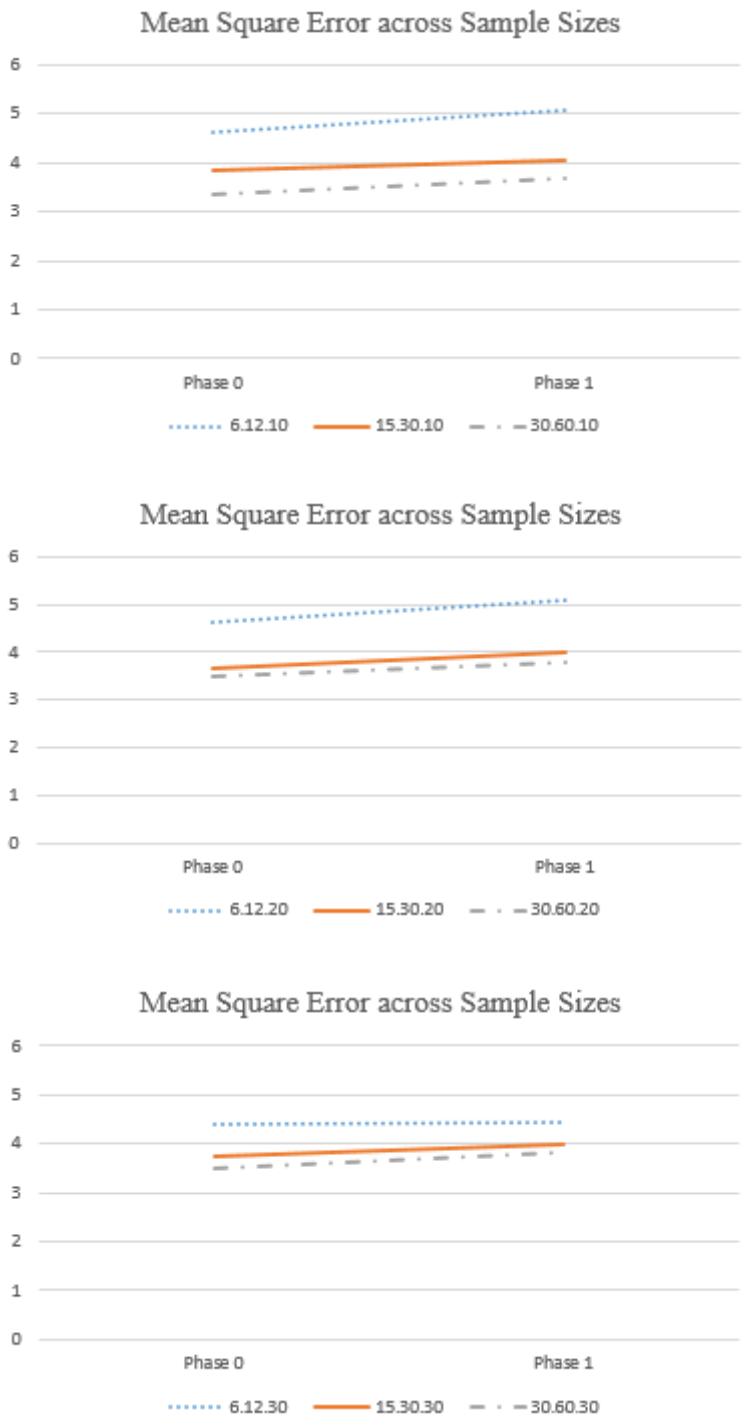

Figure 5. Mean square error across sample sizes and time-series lengths for the normal distribution. The normal distribution produced substantially higher levels of mean square error. As sample size increases, the mean square error levels decrease slightly. The time-series length of 30 observation produced the most stable levels of mean square error. See *Table 2* for key to interpret figures.

Poisson Distribution

The Poisson distribution produced substantially lower levels of mean square error than the normal distribution (e.g., $MSE \approx 0.832$ versus $MSE \approx 5.09$, respectively). The results of the Poisson distribution indicated that the level of mean square error varied based on the sample size and time-series length (see *Figure 6*). The smallest sample size ($j = 12$, $k = 6$) produced the highest levels of mean square error in the baseline phase ($min = 0.705$, $max = 0.832$) and in the intervention phase ($min = 0.458$, $max = 0.576$) across the three time-series lengths. The largest sample size ($j = 60$, $k = 30$) produced the lowest levels of mean square error in the baseline phase ($min = 0.101$, $max = 0.136$) and in the intervention phase ($min = 0.07$, $max = 0.097$) across the three-time series lengths. As the sample size increased, there was a decrease in the mean square error across time-series lengths.

The time-series lengths of 10, 20, and 30 observations produced relatively consistent levels of mean square error across the three sample size conditions (e.g., $MSE \approx 0.26 – 0.32$, $MSE \approx 0.135 – 0.10$1). The increase in time-series length produced levels of mean square that were lower across the three sample sizes, with the time-series length of 30 observations producing the lowest levels of mean square error, especially for the smallest sample size.

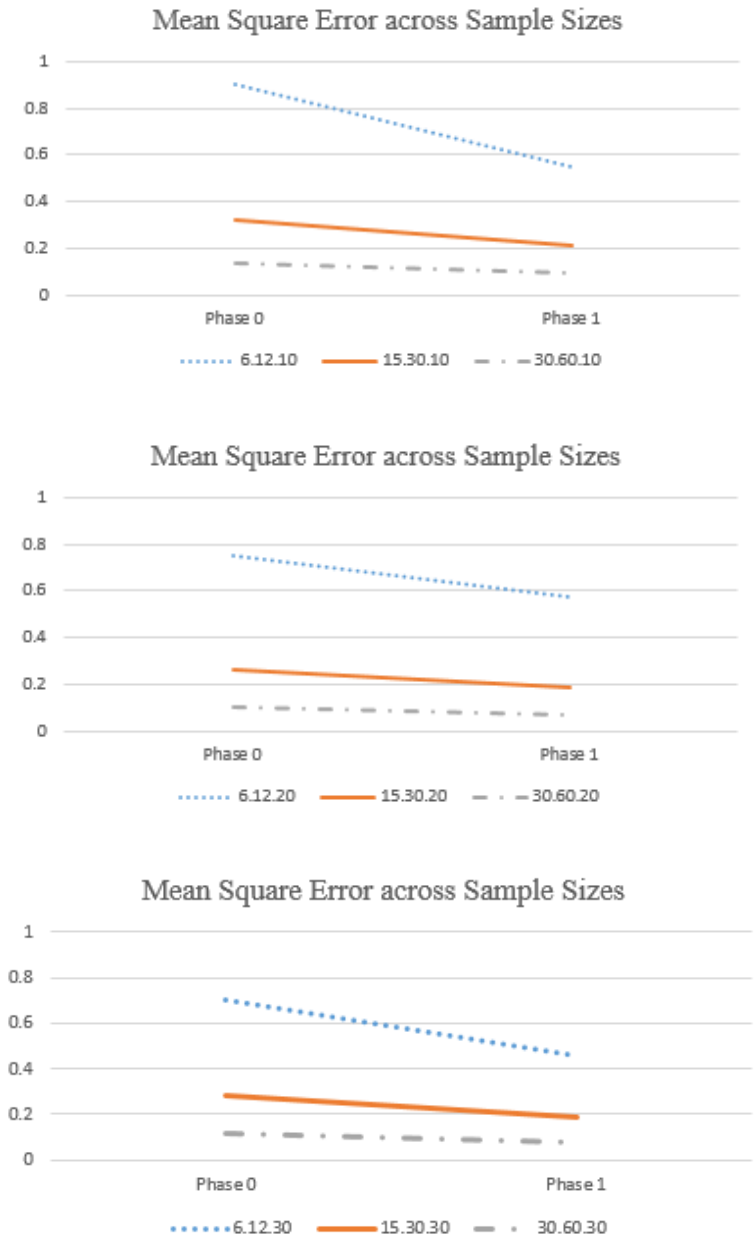

Figure 6. Mean square error across sample sizes and time-series lengths for Poisson distribution. The smallest sample size produced the highest levels of mean square error across time-series lengths. The time-series length had more of an impact on mean square error for the smallest sample size, with minor fluctuations in the other two sample sizes. The increase in time-series length results in a decrease the fluctuations between sample sizes. See *Table 2* for key to interpret figures.

Negative Binomial Distribution

The Negative Binomial distribution produced substantially lower levels of mean square error than the normal distribution (*MSE* ≈ 0.002 versus *MSE* ≈ 5.09, respectively), and the Poisson distribution (*MSE* ≈ 0.002 versus *MSE* ≈ 0.832, respectively). See *Figure 8*. The results of the Negative Binomial distribution indicated that the level of mean square error was influenced by more by the sample size than the time-series length (see *Figure 7*). The largest sample size (*j = 60, k = 30*) produced the lowest levels of mean square error across time-series lengths in the baseline phase (*min = 0.002, max = 0.022*) and in the intervention phase (*min = 0.003, max = 0.003*). The smallest sample size (*j = 12, k = 6*) produced the highest levels of mean square error across the three time-series lengths in the baseline phase (*min = 0.016, max = 0.0165*) with increases in the intervention phase (*min = 0.022, max = 0.024*). Overall, the increase in sample size produced a decrease in mean square error levels. The time-series lengths of 10, 20, and 30 observations produced minimal differences across sample sizes.

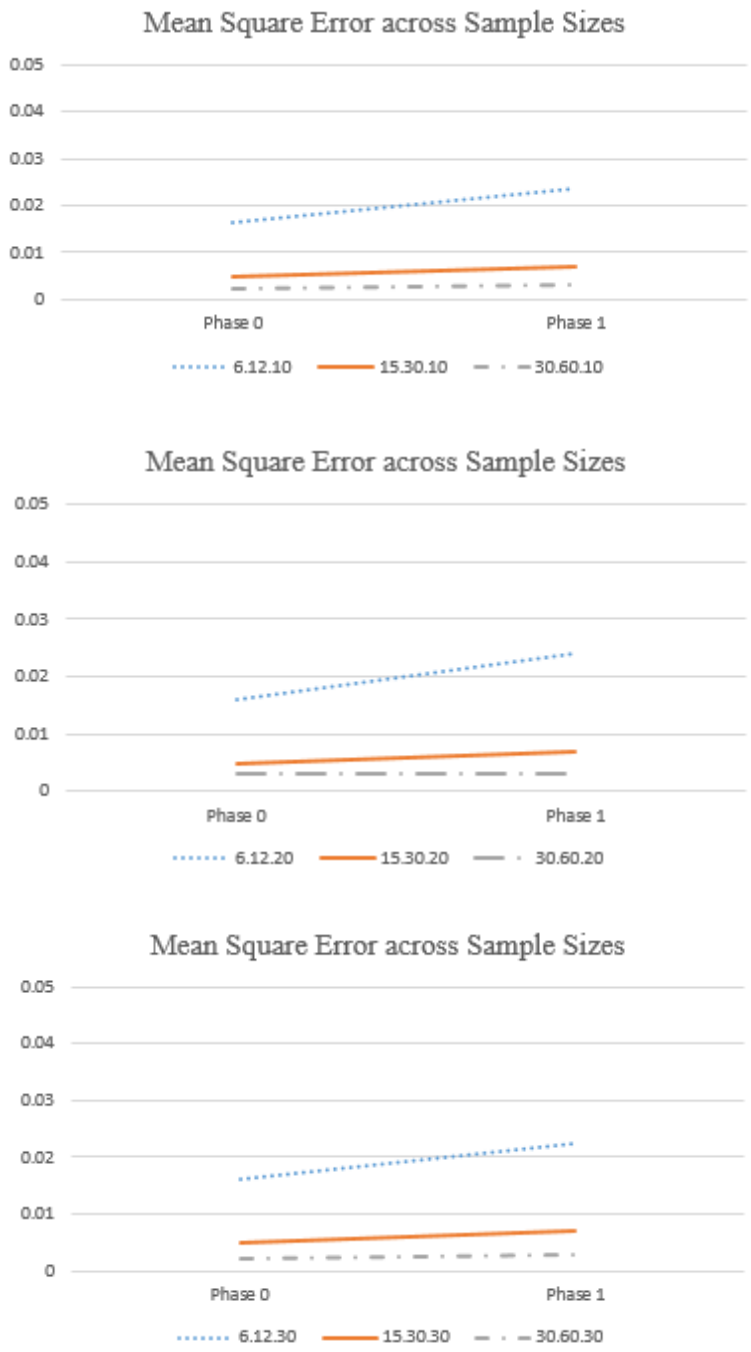

Figure 7. Mean square error across sample sizes and time-series lengths for the Negative Binomial distribution. The largest sample size produced the lowest levels of mean square error, with the smallest sample size producing the highest levels of mean square error. The three time-series lengths produced relatively consistent levels of mean square error. See *Table 2* for key to interpret figures.

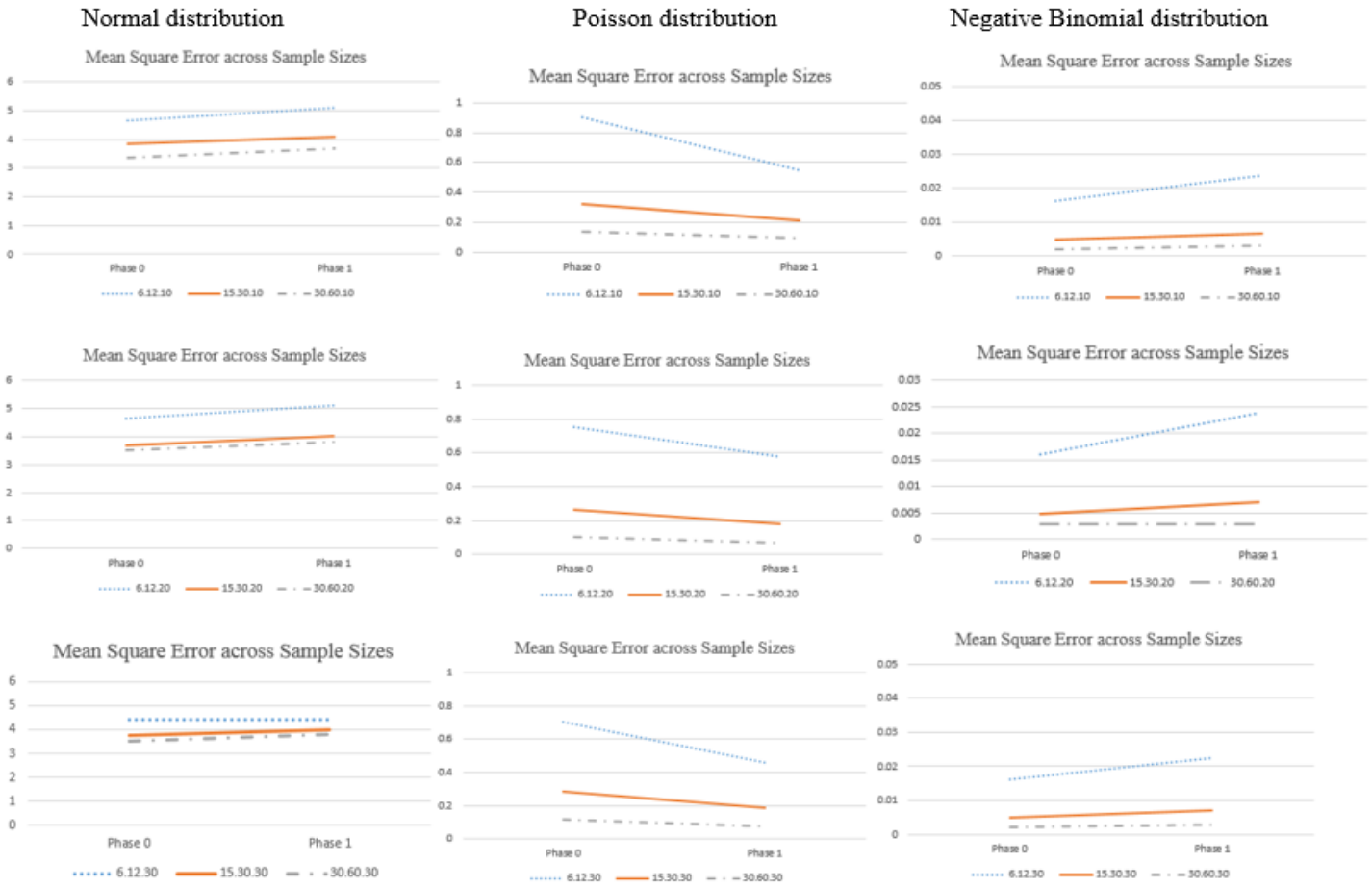

Figure 8. Comparison of mean square error across distributional assumptions, sample size, and time-series lengths (10, 20, and 30 observations, respectively). Note: the y-axes are scaled differently in order to display coverage levels clearly.

Coverage across Theoretical Distribution, Dependent on Sample Size, Degree of Freedom, and Time-Series Length

Normal Distribution

The confidence interval coverage rates for the normal distribution tended to undercover across sample sizes and time-series lengths, such that the smallest sample size (*j = 12, k = 6*) produces higher coverage levels, with coverage ranging from 0.845 to 0.642, across all three time-series lengths. The second sample size (*j = 30, k = 15*) and the largest sample size (*j = 60, k = 30*) tended to severely undercover across time-series lengths in the baseline phase (*min = 0.001, max = 0.165*) in the intervention phase (*min = 0.001, max = 0.218*).

The coverage levels were relatively consistent and tended to undercover across time-series lengths for the degree of freedom methods implemented (see *Figures 9 – 11*). The between-within and residual methods produced lower coverage levels in the baseline phase (*min = 0.642, max = 0.645*) and in the intervention phase (*min = 0.72, max = 0.759*) for the smallest sample size. The between-within and residual methods produced substantially lower coverage levels in the baseline phase (*min = 0.12, max = 0.118*) and in the intervention phase (*min = 0.159, max = 0.175*) for the second sample size. The between-within and residual methods produced lower coverage levels in the baseline phase (*min = 0.001, max = 0.002*) and in the intervention phase (*min = 0, max = 0.001*) for the largest sample size. The containment degree of freedom method resulted in coverage levels of zero across all sample size conditions and time-series lengths.

The Kenward-Roger and Satterthwaite methods produced higher coverage levels than the other degree of freedom methods across time-series lengths. For the smallest sample size, the confidence interval coverage rates for the Kenward-Roger and Satterthwaite methods tended to

undercover in the baseline phase (*min = 0.785, max = 0.788*) and in the intervention phase (*min = 0.831, max = 0.851*). For the second sample size, the Kenward-Roger and Satterthwaite methods produced lower coverage rates in the baseline phase (*min = 0.162, max = 0.18*) and intervention phase (*min = 0.204, max = 0.232*). For the largest sample size, the Kenward-Roger and Satterthwaite methods produced lower coverage rates in the baseline phase (*min = 0.001, max = 0.002*) and in the intervention phase (*min = 0.002, max = 0.003*).

The interaction between the largest sample size and time-series lengths produced the most variability among coverage levels (see *Figures 9 – 11*). The shortest time-series length of 10 observations produced coverage levels that increased from baseline to intervention phase across the degree of freedom methods. The second time-series length produced coverage levels that decreased from baseline to intervention phase for the between-within and residual methods. It appears that the Kenward-Roger and Satterthwaite methods display "flatter" trend as sample size increases across the baseline to intervention phase. The largest time-series length produced coverage levels that decreased from baseline to intervention phase for the Kenward-Roger and Satterthwaite methods. There appears to a pattern inconsistency for the between-within and residual, in which there is a minor increase of 0.001 from baseline to intervention. This inconsistency may be due to the inadequate performance of the normal distribution when recurrent event (count) outcomes are present within the data.

Summary: The normal distribution tended to undercover across the sample sizes and time-series lengths, with more severe undercoverage occurring as sample size increased (See *Figure 12*). All degree of freedom methods tended to undercover across sample sizes and time-series lengths. Overall, the normal distribution performed relatively better in terms of confidence interval coverage levels when sample size was smallest, dependent on time-series

length. The normal distribution also demonstrated better coverage levels for the Kenward-Roger and Satterthwaite degree of freedom methods for the smallest sample size. These results suggest that the normal distribution performs worse in research situations where sample size is larger.

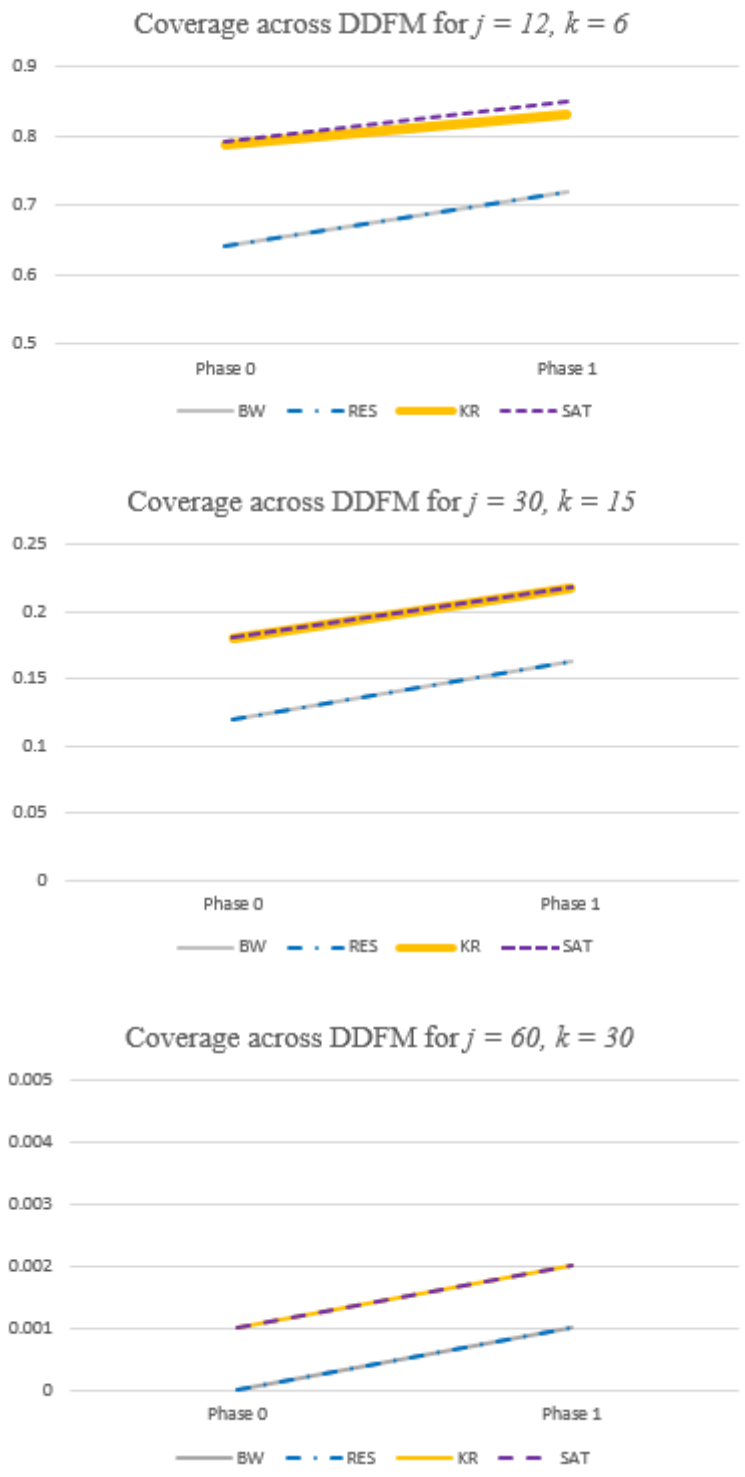

Figure 9. Coverage across degree of freedom methods across sample size for a time-series length of 10 observations for normal distribution. The confidence interval coverage rates tended to undercover across degree of freedom methods, with Kenward-Roger and Satterthwaite methods producing higher coverage rates. The coverage rates were highest for the smallest sample size.

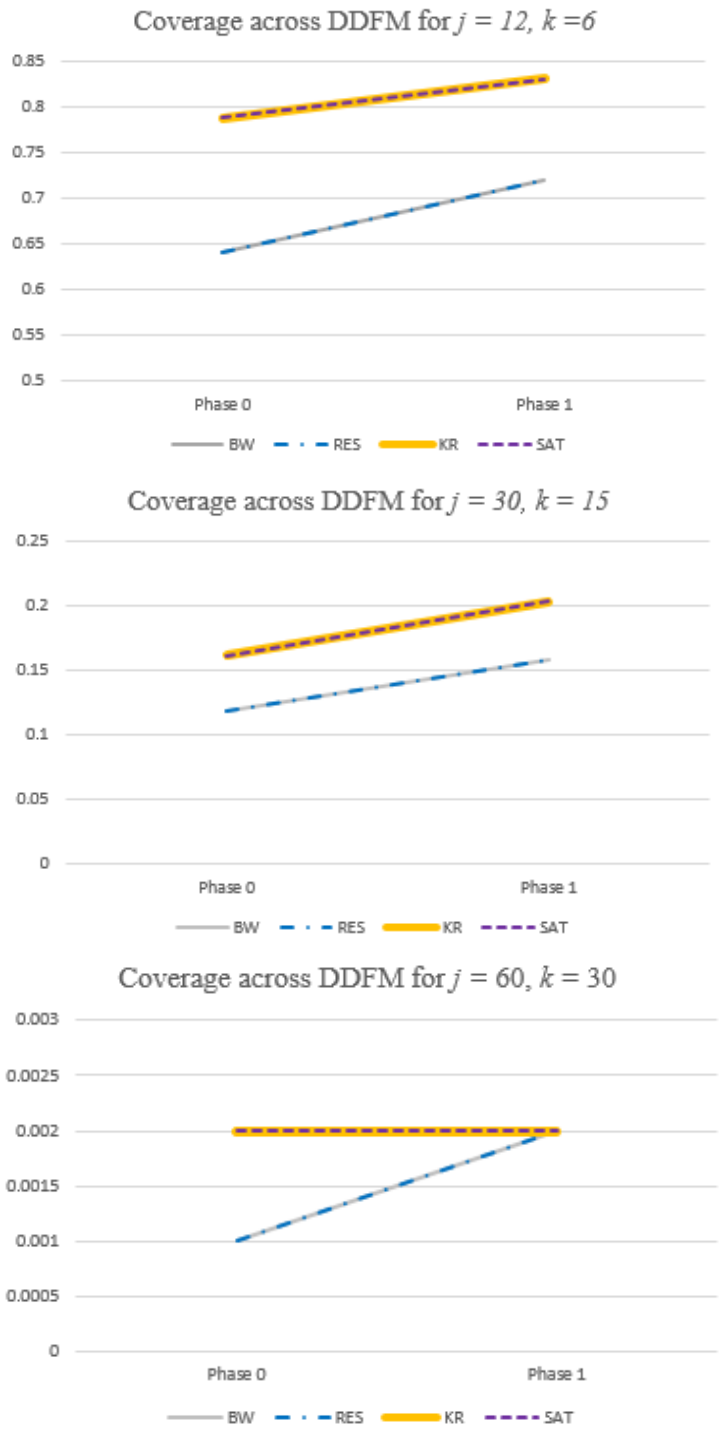

Figure 10. Coverage across degree of freedom methods across sample size for 20 observations for normal distribution. The confidence interval coverage rates tended to undercover across degree of freedom methods, with Kenward-Roger and Satterthwaite methods producing higher coverage rates. The coverage rates were highest for the smallest sample size. Note: the residual and between-within degree of freedom methods increased minimally (0.001 to 0.002) for the largest sample size condition. Due to scaling requirements, the increase appears more dramatic than in actuality.

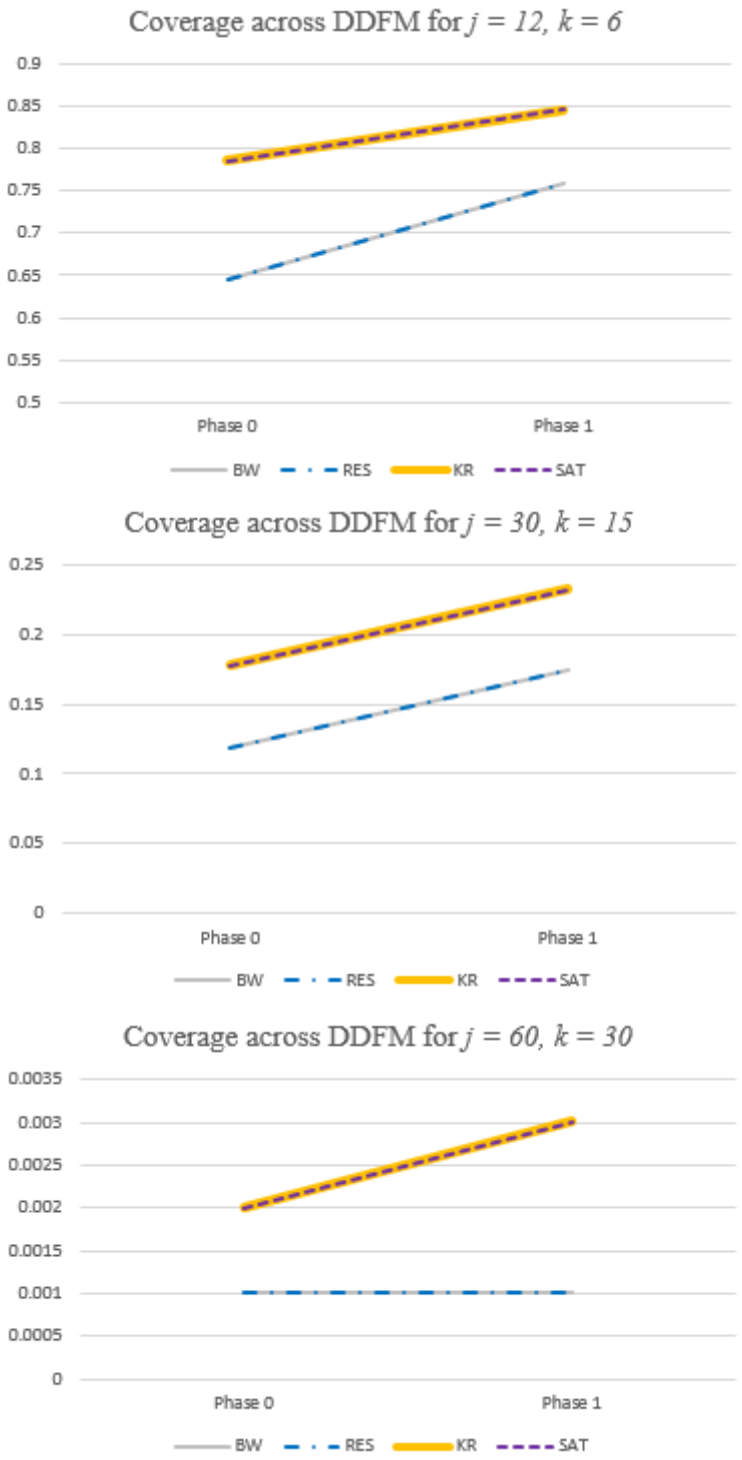

Figure 11. Coverage across degree of freedom methods across sample size for 30 observations for normal distribution. The confidence interval coverage rates tended to undercover across degree of freedom methods, with Kenward-Roger and Satterthwaite methods producing higher coverage rates. The coverage rates were highest for the smallest sample size. Note: There is a pattern inconsistency for the residual and between-within degree of freedom method, in which the coverage rate does not change from baseline to intervention.

Figure 12. Comparison of confidence interval coverage rates across degrees of freedom methods, sample size, and the time-series lengths (10, 20, and 30 observations, respectively) for the normal distribution.

Poisson Distribution

The Poisson distribution produced confidence interval coverage rates that were closer to the nominal .95 level across sample sizes and time-series lengths, with variation in coverage across degree of freedom methods (see *Figures 13 – 15).* For the smallest sample size (*j = 12, k = 6*), the between-within and residual methods tended to undercover in the baseline phase (*min = 0.898, max = 0.928*) and in the intervention phase (*min = 0.915, max = 0.943*) across time-series lengths. For the second sample size (j = 30, k = 15), the between-within and residual methods tended to undercover in the baseline phase (*min = 0.898, max = 0.912*) and in the intervention phase (*min = 0.899, max = 0.927*) across time-series lengths. For the largest sample size (*j = 60, k = 30*), the between-within and residual methods tended to undercover in the baseline phase (*min = 0.911, max = 0.925*) and in the intervention phase (*min = 0.882, max = 0.94*) across the time-series lengths.

For the smallest sample size, the Kenward-Roger and Satterthwaite degree of freedom methods tended to overcover in the baseline phase (*min = 0.959, max = 0.964*) and in the intervention phase (*min = 0.943, max = 0.972*) for the second and largest time-series lengths, whereas the methods tended to undercover in the baseline phase (0.941) and intervention phase (0.955) for the smallest time-series lengths. For the second sample size, the Kenward-Roger and Satterthwaite methods tended to undercover in the baseline phase (*min = 0.924, max = 0.947*) and in the intervention phase (*min = 0.922, max = 0.948*) across the time-series lengths. For the largest sample size, the Kenward-Roger and Satterthwaite methods tended to undercover in the baseline phase (*min = 0.924, max = 0.952*) and in the intervention phase (*min = 0.907, max = 0.956*) across the time-series lengths.

The shortest time-series length of 10 observations produced higher coverage rates for the Kenward-Roger and Satterthwaite degree of freedom methods in the baseline phase (*min = 0.923, max = 0.941*) and in the intervention phase (*min = 0.907, max = 0.955*), compared to the between-within and residual methods in the baseline phase (*min = 0.898, max = 0.924*) and in the intervention phase (*min = 0.882, max = 0.924*) across sample sizes. The second time-series length of 20 observations produced higher coverage rates for the Kenward-Roger and Satterthwaite methods in the baseline phase (*min = 0.929, max = 0.959*) and in the intervention phase (*min = 0.927, max = 0.946*) across sample sizes. The largest time-series length of 30 observations produced higher coverage rates for the Kenward-Roger and Satterthwaite degree of freedom methods in the baseline phase (*min = 0.926, max = 0.964*) and in the intervention phase (*min = 0.94, max = 0.972*), compared to the between-within and residual methods.

Summary: The Poisson distribution performed better in terms of coverage rates than the normal distribution, with the coverage closer to the nominal .95 level and the tendency to undercover (See *Figure 16*). The increase in sample size produced minor decreases in coverage levels, and the increase in time-series lengths resulted in no discernable fluctuation in coverage levels. The between-within and residual degree of freedom methods produced the lowest coverage levels than the Kenward-Roger and Satterthwaite methods. The most severe overcoverage occurred when the sample size was smallest in the shortest time-series length for the Kenward-Roger and Satterthwaite methods.

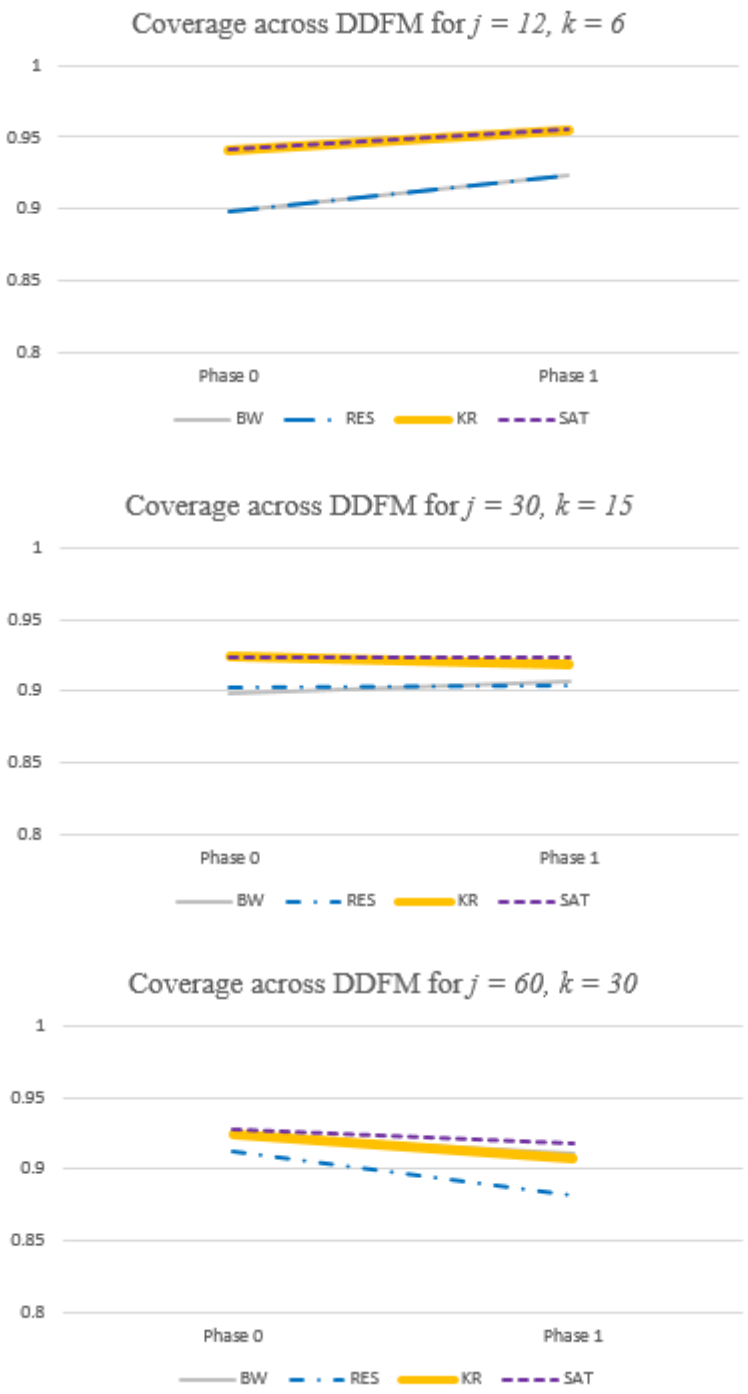

Figure 13. Coverage across degree of freedom methods across sample size for 10 observations for Poisson distribution. The confidence interval coverage rates tended to undercover across the degree of freedom methods and sample sizes, but with coverage levels closer to the nominal .95 level than the normal distribution. The coverage rates were highest for the Kenward-Roger and Satterthwaite methods. There were decreases in coverage when sample size increased.

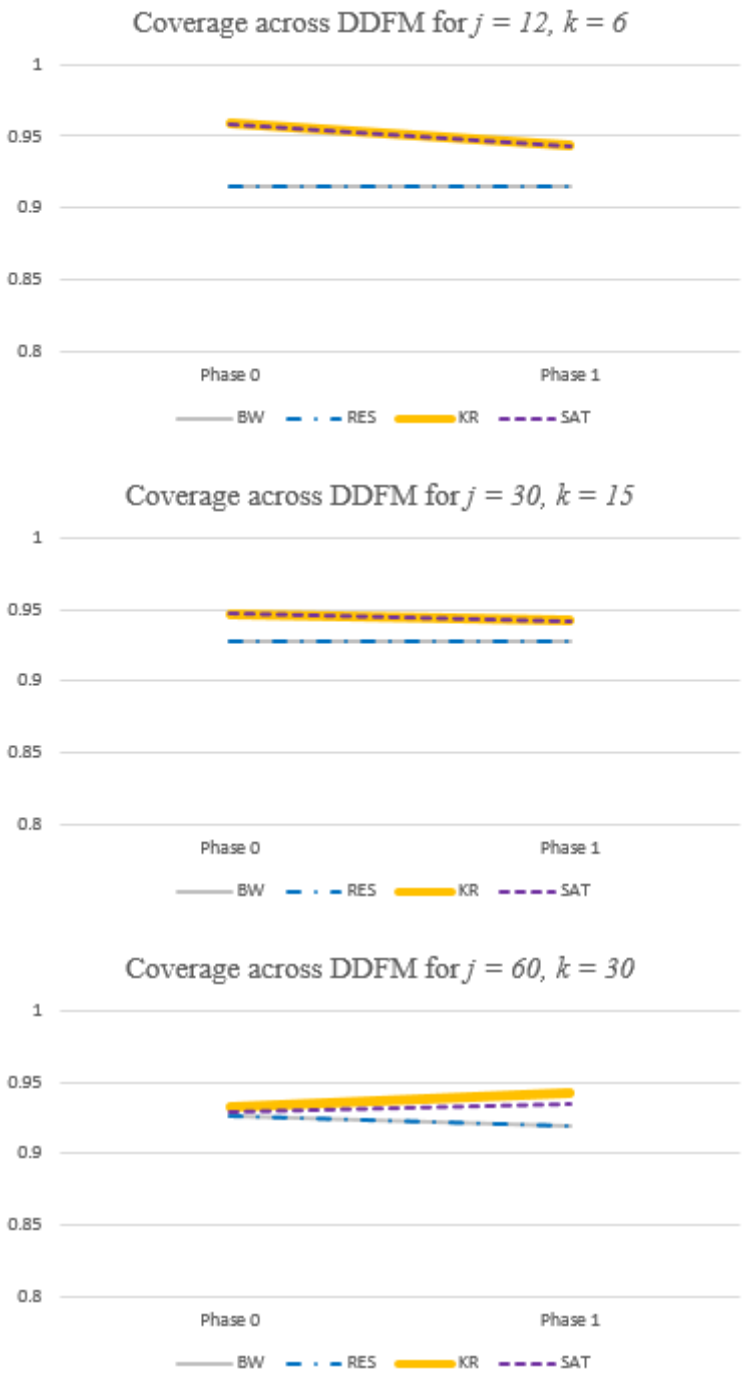

Figure 14. Coverage across degree of freedom methods across sample size for 20 observations for Poisson distribution. The confidence interval coverage rates tended to undercover across the degree of freedom methods and sample sizes, but with coverage levels closer to the nominal .95 level than the normal distribution. The coverage rates were highest for the Kenward-Roger and Satterthwaite methods. There were minor decreases in coverage as sample size increased.

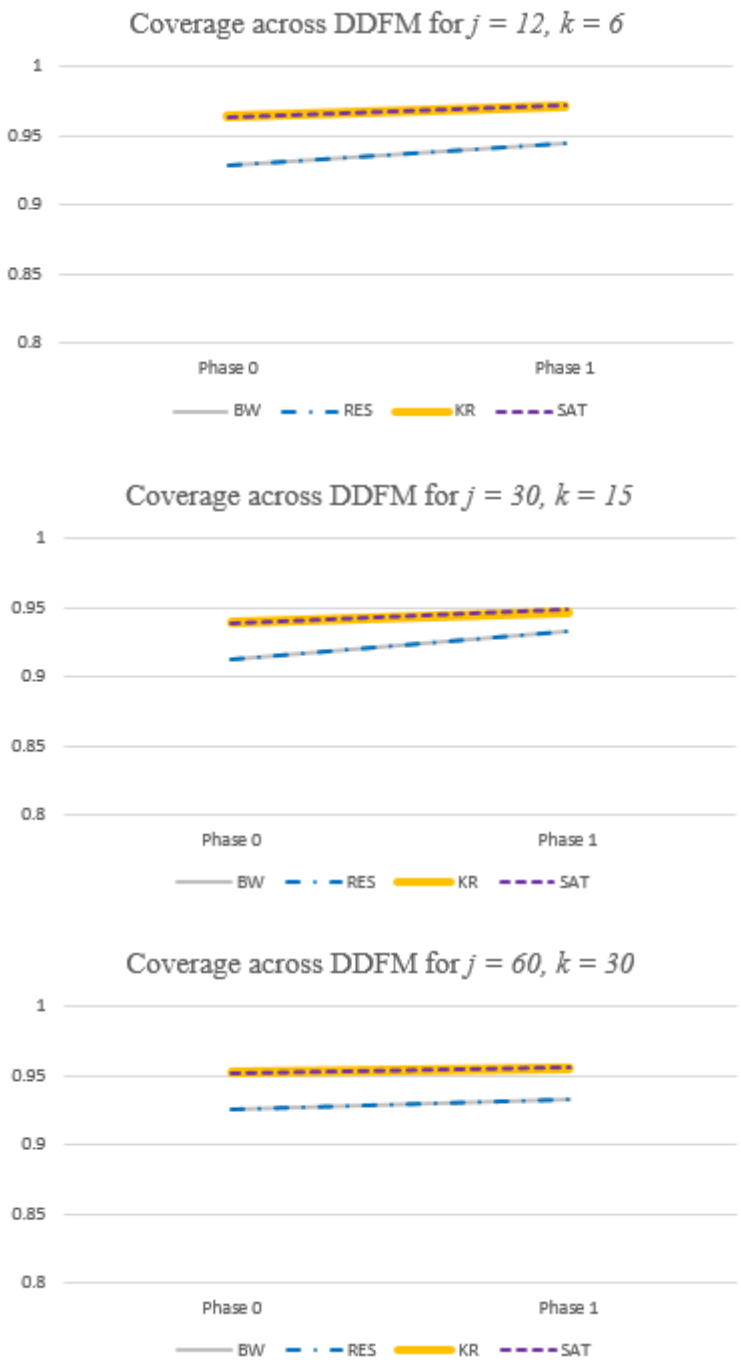

Figure 15. Coverage across degree of freedom methods across sample size for 30 observations for Poisson distribution. The confidence interval coverage rates tended to undercover across the degree of freedom methods and sample sizes, but with coverage levels closer to the nominal .95 level than the normal distribution. The coverage rates were highest for the Kenward-Roger and Satterthwaite methods. There were minor fluctuations in coverage as sample size increased.

Figure 16. Comparison of confidence interval coverage rates across degrees of freedom methods, sample size, and time-series lengths (10, 20, and 30 observations, respectively) for the Poisson distribution.

Negative Binomial Distribution

The Negative Binomial distribution produced confidence interval rates that were closer to the nominal .95 level across sample sizes and time-series lengths, with differences occurring in coverage across degree of freedom methods (see *Figures 17 – 19*). For the smallest sample size (*j = 12, k = 6*), the between-within and residual methods tended to undercover in the baseline phase (*min = 0.903, max = 0.91*) with minor increases in the intervention phase (*min = 0.907, max = 0.918*), independent of time-series lengths. For the second sample size (*j = 30, k = 15*), the between-within and residual methods also tended to undercover in the baseline phase (*min = 0.929, max = 0.939*) and in the intervention phase (*min = 0.935, max = 0.941*). For the largest sample size (*j = 60, k = 30*), the between-within and residual methods tended to undercover in the baseline phase (*min = 0.939, max = 0.94*) and tended to minimally overcover in the intervention phase (*min = 0.942, max = 0.952*).

For the three sample size conditions, the Kenward-Roger and Satterthwaite methods generally tended to fluctuate in overcoverage and undercoverage in the baseline phase (*min = 0.941, max = 0.955*) and in the intervention phase (*min = 0.941, max = 0.955*) across time-series lengths. For the smallest sample size, the Kenward-Roger and Satterthwaite methods tended to decrease in coverage as time-series length increased for the baseline phase (*max = 0.955, min = 0.941*) and for the intervention phase (*max = 0.955, min = 0.941*). For the second sample size, the two methods tended to produce minor increases in coverage as time-series length increased for the baseline phase (*min = 0.946, max = 0.959*) and for the intervention phase (*min = 0.947, max = 0.952*). For the largest sample size, the two methods tended to produce minor increases in coverage as time-series length increased for the baseline phase (*min = 0.944, max = 0.949*) and for the intervention phase (*min = 0.948, max = 0.956*).

The shortest time-series length of 10 observations produced minor decreases in coverage rates as the sample size increased for the baseline phase (*min = 0.944, max = 0.951*) and for the intervention phase (*min = 0.947, max = 0.955*). For the second time-series length of 20 observations, the confidence interval coverage rates were relatively consistent as sample size increased for the baseline phase (*min = 0.949, max = 0.955*) and for the intervention phase (*min = 0.945, max = 0.956*). For the largest time-series length of 30 observations, the coverage rates produced relatively consistent coverage rates as the sample size increased for the baseline phase (*min = 0.941, max = 0.941*) and for the intervention phase (*min = 0.959, max = 0.952*).

Summary: The Negative Binomial distribution performed better in regards to the coverage rates than the normal distribution and the Poisson distribution, with coverage rates generally closer to the nominal .95 level (See *Figure 20*). The increase in sample size resulted in relatively consistent levels of coverage rates, and the increase in time-series lengths generally produced an increase in coverage rates across sample sizes. The most severe undercoverage occurred when the sample size was largest with the smallest time-series length for the Kenward-Roger and Satterthwaite methods. The most severe overcoverage occurred within the second sample size and the largest time-series length for the Kenward-Roger and Satterthwaite methods.

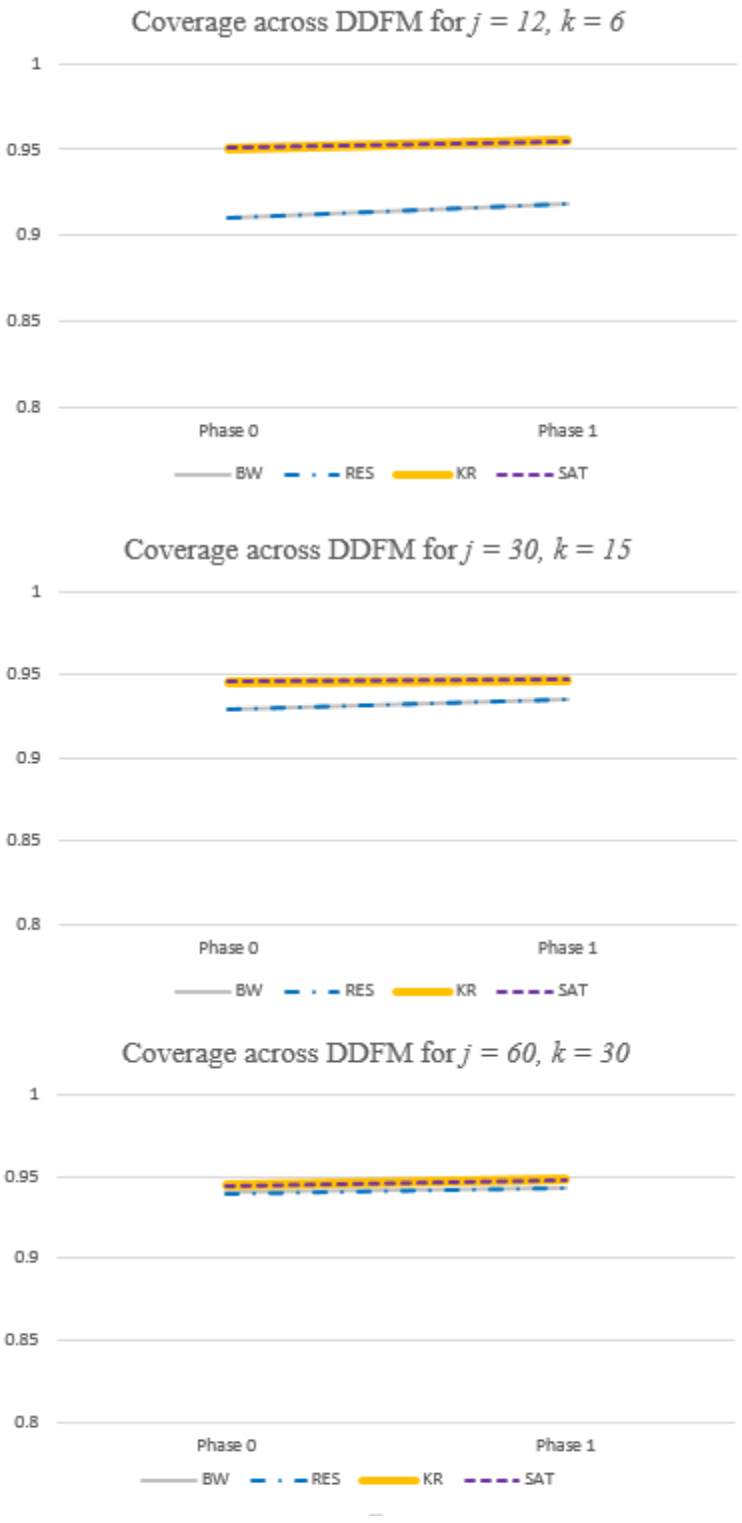

Figure 17. Coverage across degree of freedom methods across sample size for 10 observations for the Negative Binomial distribution. There were less discernable fluctuations between degree of freedom methods as sample size increased. Generally speaking, the degrees of freedom methods tended to undercover.

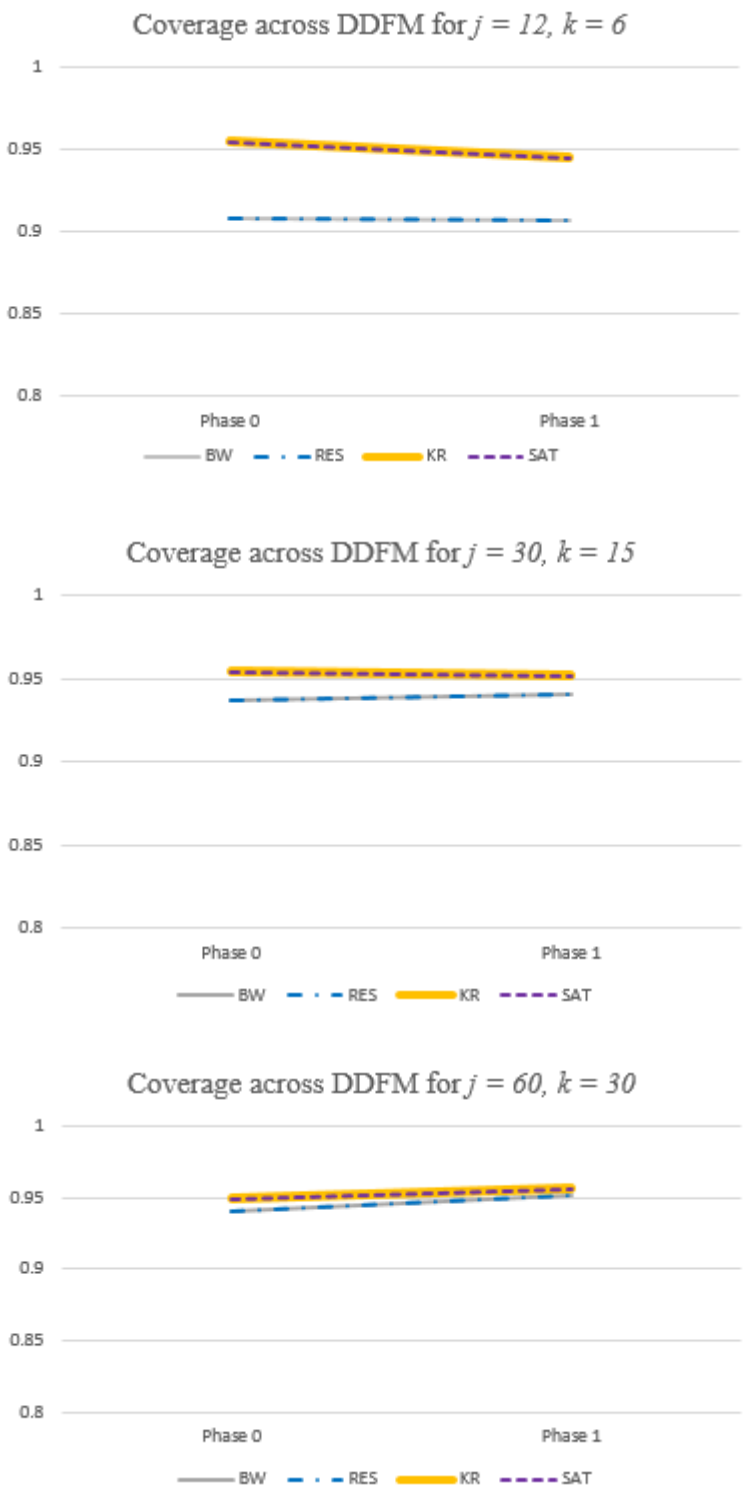

Figure 18. Coverage across degree of freedom methods across sample size for 20 observations for the Negative Binomial distribution. There were less discernable fluctuations between degree of freedom methods as sample size increased. Overcoverage occurred more frequently in the Negative Binomial distribution than the other two distributions.

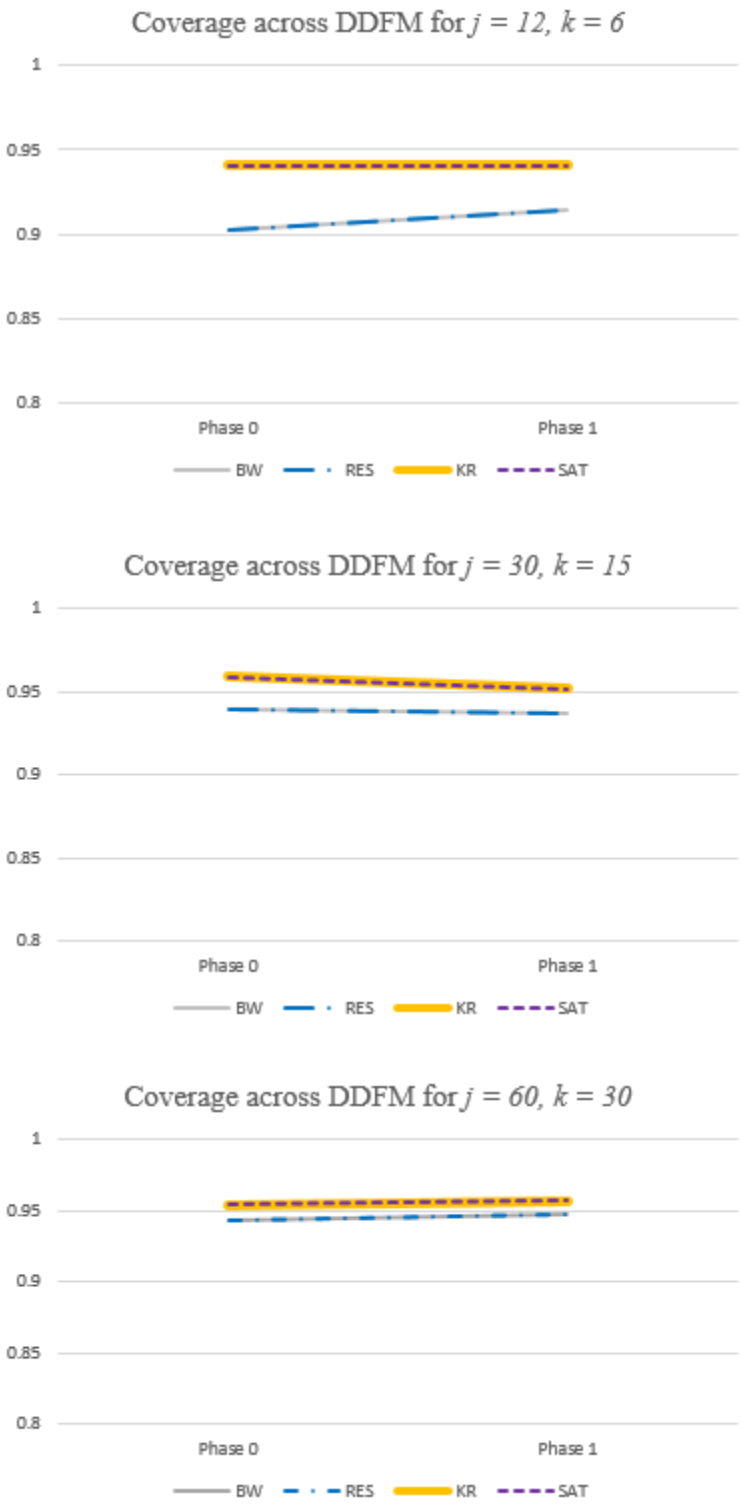

Figure 19. Coverage across degree of freedom methods across sample sizes for 30 observations for the Negative Binomial distribution. There were less discernable discrepancies between degree of freedom methods as sample size increased. Overcoverage occurred more frequently in the Negative Binomial distribution than the other two distributions.

Figure 20. Comparison of confidence interval coverage rates across degrees of freedom methods, sample size, and time-series lengths (10, 20, and 30 observations, respectively) in the Negative Binomial distribution.

Discussion

The recent literature on synthesized single-case designs has provided significant contributions to the understanding of meta-analysis techniques for single-case design data (Owens, 2011; Ugille et al., 2012; Van den Noortgate & Onghena, 2003, 2008). The application of meta-analysis to SCD data means that researchers are no longer limited to the study of either groups or individuals; all pieces of information can be obtained. The advancement of synthesized single-case design data also allows researchers to move beyond traditional visual inspection (e.g., Kratochwill & Levin, 2014) and towards more sophisticated modeling techniques, like multilevel modeling (Owens, 2011; Rindskopf & Ferron, 2014; Van den Noortgate & Onghena, 2008). The benefits of implementing multilevel within the context of synthesized single-case design data have been documented, which include the ability to account for the hierarchical structure of repeated observations nested within individuals, and with individuals nested within higher levels (e.g., schools or communities; Gill & Womack, 2013), the ability to account for interdependency due to repeated observations (Baek et al., 2014), and the flexibility to adapt a given model to the unique specifications of the SCD data (Van den Noortgate & Onghena, 2003a).

The purpose of this study was to evaluate the utility of multilevel modeling through meta-analytic techniques for handling recurrent event (count) outcomes in single-case designs. A combination of empirical and Monte Carlo simulation methods were used. The results of the current study are based on the three-level theoretical model from stage one of the research design. Results were compared across different distributional assumptions, different sample sizes, different time-series lengths, and different degree of freedom methods based on levels of relative bias, mean square error, and confidence interval coverage rates.

Specifically for the primary purpose of the current research, the combination of multilevel modeling and meta-analytic techniques allows researchers to synthesize single-case design data for the purposes of observing and analyzing the "population" effects from the sample, rather than the "intra-individual" effects alone. The use of multilevel modeling further allows single-case design researchers to account for the non-normality involved in recurrent event (count) outcomes, which are prevalent in SCD research (Shadish & Sullivan, 2011). The ability to modify the distributional assumption from normal to either Poisson or Negative Binomial, in order to reflect the true nature of the data, provides additional support for the use of multilevel modeling in the context of synthesized SCDs where recurrent event (count) outcomes are present. Multilevel modeling also provides single-case design researchers with the opportunity to account for the hierarchical structure present in SCD data when repeated observations (level-1) are nested within students (level-2), and students are nested within teachers (level-3; Gill & Womack, 2013).

The results indicate that the distributional assumption should reflect the true distribution of the data set. The Negative Binomial distribution outperformed the normal distribution across sample sizes and time-series lengths in terms of relative bias, mean square error, and coverage rates. More specifically, the Negative Binomial distribution produced lower levels of relative bias, lower levels of mean square error, and better confidence interval coverage rates (~ .95 nominal level). The Poisson distribution also outperformed the normal distribution across sample sizes and time-series lengths, with lower levels of relative bias and mean square error, as well as better confidence interval coverage rates. The Poisson distribution performed similarly to the Negative Binomial distribution across sample sizes and time-series lengths, with the Negative Binomial distribution producing lower levels of relative bias and mean square error. In

short, the Negative Binomial distribution outperformed the other two distributional assumptions in this research study. The subsequent discussion, therefore, discusses the remaining conditions (i.e., sample size, time-series lengths, and degree of freedom methods) based on the Negative Binomial distribution.

As hypothesized, the Kenward-Roger and Satterthwaite degree of freedom methods performed closer to the .95 nominal level for confidence interval coverage rates than the between-within, residual, and containment methods across sample sizes and time-series lengths. The prevalence of overcoverage occurred, on average, in conditions where the sample size was larger and the time-series length was longer. The between-within and residual methods consistently undercovered, which is consistent with previous research (Ferron et al., 2009). The discrepancies in coverage rates between degree of freedom methods were minimized with the increase in sample size and time-series length for the Negative Binomial distribution. The discussion for degree of freedom methods focus on the coverage rates, because the relative bias and mean square error levels are unaffected by varying the degree of freedom method.

The variation in time-series length did not produce considerable differences in regards to relative bias, mean square error, and coverage rates when the true distributional assumption is selected. More specifically, the decisions for distributional assumption and degree of freedom method are considered more influential in the final study outcomes. The increase in time-series length, on average, minimally decreased relative bias levels and mean square error levels as well as increased coverage rates in the Negative Binomial distribution. The variation in sample size produced differences had more of an influence on mean square error, relative bias, and coverage rates when the true distributional assumption is selected. The increase in sample size increased coverage rates, further decreased relative bias levels, and further decreased mean square error

levels for the Negative Binomial distribution. More specifically, the largest sample size produced the lowest levels of mean square error and relative bias, with better coverage rates.

Recommendations for Future Research

The long-term goal of this study is to provide recommendations and guidelines regarding the appropriate methodological decisions/approaches for handling count outcomes in synthesized single-case design data within the same study. Therefore, the recommendations are as follows:

1. The presence of count data should necessitate the use of a distribution that will account for the non-normality, such as Poisson or Negative Binomial (Shadish et al., 2013). The substantially lower levels of relative bias and mean square error, as well as acceptable confidence interval coverage rates, provides support that the Negative Binomial outperformed the other two distributions, with the Poisson distribution performing substantially better than the normal distribution.
2. The Kenward-Roger and Satterthwaite degree of freedom methods are recommended for studies with various sample sizes, especially for smaller sample sizes. When cluster sizes are balanced, then Kenward-Roger and Satterthwaite will perform similarly. Based on the results of this research, Kenward-Roger and Satterthwaite produced coverage rates that were, on average, closer to the nominal level .95 for both the Poisson and Negative Binomial distribution. This recommendation is consistent with Ferron et al. (2009).
3. The length of the time series should be based on the number of observations required to establish experimental control, while considering the WWC standards (Kratchowill et al., 2010). The results of the current research found minor differences when the time-series length was varied in terms of relative bias, mean square error, and

confidence interval coverage rates; therefore, other factors (i.e., distributional assumption and degree of freedom methods) were more influential.

4. The sample size of level-two and level-three should be increased when possible to produce better estimates. Findings from previous research recommend a minimum sample size of 30 units at the upper level (Hox, 1998). The results of the current research found that the smallest sample size consistently produced the highest levels of mean square error and relative bias, with the largest sample size producing the lowest levels of mean square error and relative bias. The increase in sample size length produced less variability in coverage rates between degrees of freedom methods for the Poisson and Negative Binomial distributions.

Strengths of the Current Study

First, the current study accounted for and compared non-normal distributions (i.e., Poisson and Negative Binomial) in the context of synthesized single-case designs where the dependent variable is a recurrent event (count) outcome. The ability to account for non-normal distributions is important, especially within the single-case design framework. Second, the current study utilized real-world SCD empirical data to generate single-case design data for several sample sizes and time-series lengths for the purposes of comparing the relative benefits of varying each condition. The use of real SCD data to inform Monte Carlo simulations can strengthen the utility of data generation. Third, the current study provided additional empirical support for the recommendation of the Kenward-Roger and Satterthwaite degree of freedom methods in synthesized single-case design research, as stated in Ferron et al. (2009).

Limitations of the Current Study

There were several methodological decisions in the current study that, while justified based on the current purpose, should be noted as limitations of the current findings. First, the two-level data-driven model from the empirical CBC data deviated from the three-level theoretical model. This decision was made to avoid modifying the multilevel model as a direct result of an empirical data set, in which there could be no theoretical justification for said model (see *Justification* on page 48). Second, the use of the CBC RCTs represents an unconventional direction for SCD data; however, the CBC data still represent a multiple baseline design across individuals. The SCD data were then synthesized for the purposes of the current study. The RCTs simply provided a broader context for the single-case design framework.

Finally, the synthesized empirical data set contained the control and treatment participants for the purposes of increasing the level-three sample size. The drawback to this decision was that intervention effects could not be assessed, because the presence of the control participants counteracted any intervention effect. The purpose of the current study was not to assess the treatment effects across time, but instead to assess the utility of multilevel modeling when recurrent event (count) outcomes are present for synthesized single-case designs. Additionally, previous research has provided empirical support for increasing the level-3 sample size in order to achieve greater accuracy (Owens & Ferron, 2012). That said, traditional SCD research are interested in assessing the possible intervention effects, and it should be noted that the methodological decisions of this study resulted in a failure to assess these effects.

Future Directions

First, the current study utilized a "short time-series" approach to single-case design research for assessing recurrent event (count) outcomes. An extension of this research could benefit from increasing the number of observations within each phase to represent a true time-

series design, in which there are 50 – 100 observations within the baseline phase and the intervention phase (Greene, 2000). Second, the degree of autocorrelation was also not varied or controlled for in the current study. Future research could vary the degree of autocorrelation to determine its influence on the other simulation conditions (e.g., Owens, 2011). Third, the typical SCD research study utilizes more than a simple AB design (baseline and intervention, respectively) in order to establish experimental control. For future research, including another baseline phase or perhaps another baseline and intervention phase may represent stronger experimental control and prevent threats to validity.

Finally, another distributional alternative to Poisson and Negative Binomial is the zero-inflated Poisson (ZIP) distribution, in which researchers can account for excessive "structural zeros" and "non-structural zeros" (or those that potentially have zero) in count data (Lambert, 1992). For example, children who never participate in certain disruptive behaviors can be identified based on the excessive (structural) zeros and can be removed from the data set. In this case, we would only be interested in the children who participated in these disruptive behaviors prior to the CBC intervention in order to determine the actual effect of treatment. An extension of this research may benefit from the inclusion of the zero-inflated distribution to account for the excessive zeros present in count data.

Conclusion

The findings from the current study suggest that the use of meta-analysis in combination with multilevel modeling may be a viable alternative for single-case design researchers. The results of the current study further indicate that the distributional assumption implemented within a study should reflect the true nature of the data. More specifically, when data are truly non-normally distributed, then researchers should utilize a distribution that accounts for the lack of

normality. The distributional assumption implemented within a study also appears to be the most contributing factor to the success of research studies (based on relative bias, mean square error, and confidence interval coverage rates). Researchers are advised to consider their research context, distribution of their data set, and the purpose of their research when making important methodological decisions.

References

Agresti, A. & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician, 54*, 280-288.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single-case. *Behaviour Research and Therapy, 31*, 621-631. doi: 10.1016/0005-7967(93)90115-B.

Baek, E. K., Moeyaert, M., Petit-Bois, M., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2014). The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation, 24*, 590-606. doi: 10.1080/096202011.2013.835740.

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis, 10*, 167-172.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Pearson.

Bliss, C. I., & Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics, 9*, 176-200.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234-246. doi: 10.1177/0145445503259264

Chamberlain, P., & Reid, J. B. (1987). Parent observation and report of child symptoms. *Behavioral Assessment, 9*, 97-109.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews*. London, UK: SAGE Publications Ltd.

Cooper, H. M., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*, 165-176. doi: 10.1037/a0015565.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment, 91*, 121-136. doi: 10.1080/00223890802634175.

Delgado, M. A., & Kniesner, T. J. (1997). Count data models with variance of unknown form: An application to a hedonic model of worker absenteeism. *The Review of Economics and Statistics, 79*, 41-49.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*, 372-384. doi: 10.3758/BRM.41.2.372.

Gill, J., & Womack, A. J. (2013). The multilevel model framework. In J. S. Simonoff, M. A. Scott & B. D. Marx (Eds.), *The SAGE handbook of multilevel modeling* (pp. 3-20). London, UK: SAGE Publications Ltd.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Green, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Harvey, M., May, M., & Kennedy, C. (2004). Nonconcurrent Multiple Baseline Designs and the Evaluation of Educational Systems. *Journal of Behavioral Education, 13(4),* 267-276. doi:10.1023/B:JOBE.0000044735.51022.5d

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. (unpublished manuscript).

Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147-154). New York, NY: Springer.

Hox, J., & van de Schoot, R. (2013). Robust methods for multilevel analysis. In J. S. Simonoff, M. A. Scott & B. D. Marx (Eds.), *The SAGE handbook of multilevel modeling* (pp. 387-403). London, UK: SAGE Publications Ltd.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*, 483-493. doi: 10.1002/pits.20240.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings.* (2nd ed.). New York, NY: Oxford University Press.

Kiernan, K., Tao, J., & Gibbs, P. (2012). Tips and strategies for mixed modeling with SAS/STAT procedures. SAS Global Forum 2012. Retrieved from http://support.sas.com/resources/papers/proceedings12/332-2012.pdf.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from the What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *American Statistical Association and the American Society for Quality Control, 36*, 1-12.

Lee, A. H., Wang, K., Scott, J. A., Yau, K. K. W., & McLachlan, G. J. (2006). Multi-level zero-inflated Poisson regression modeling of correlated count data with excess zeros. *Statistical Methods in Medical Research, 15*, 47-61. doi: 10.1191/0962280206sm429oa

Maas, C. J. M., & Hox, J. J. (1999). Sample sizes for multilevel modeling. (unpublished manuscript).

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92. doi: 10.1027/1614-1881.1.3.86.

Moghimbeigi, A., Eshraghian, M. R., Mohammad, K., & McArdle, B. (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics, 35*, 1193-1202. doi: 10.1080/02664760802273203

Olsson, U. (2002). *Generalized linear models: An applied approach* (pp. 31-40). Sweden: Studentlitteratur, Lund.

Owens, C. M. (2012). *Meta-analysis of single-case data: A Monte Carlo investigation of a three level model*. Retrieved from ProQuest Digital Dissertations. (AAT 3449468)

Parker, R. L., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of several statistical methods. *Behavior Therapy, 34*, 189-211.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Designs and implementation. *Structural Equation Modeling, 8*, 287-312.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*, 85-112. doi: 10.1016/j.jsp.2009.09.002.

Pohlmeier, W., & Ulrich, V. (1995). An econometric model of the two-part decision making process in the demand for health care. *The Journal of Human Resources, 30*, 339-361.

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Rindskopf, D. M., & Ferron, J. M. (2014). Using multilevel models to analyze single-case design data. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 221-246). New York, NY: Oxford University Press.

Scruggs, T. E., Mastropieri, M. A., & Castro, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24-33.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*, 385-405. doi: 10.1037/a0032964.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*. doi: 10.3758/s13428-011-0111-y.

Sheridan, S. M., Bovaird, J. A., Glover, T. A., Garbacz, S. A., Witte, A., & Kwon, K. (2012). A randomized trial examining the effects of conjoint behavioral consultation and the mediating role of the parent-teacher relationship. *School Psychology Review, 41*, 23–46.

Sheridan, S. M., Holmes, S. R., Coutts, M. J., & Smith, T. E. (2012). Preliminary effects of conjoint behavioral consultation in rural communities (R^2Ed Working Paper No. 2012-1). Retrieved from the National Center for Research on Rural Education: r^2ed.unl.edu

Sheridan, S. M., Holmes, S. R., Coutts, M. J., Smith, T. E., Kunz, G. M., & Witte, A. L. (2013). CBC in rural schools: Preliminary results of a randomized trial (R^2Ed Working Paper No. 2013-1). Retrieved from the National Center for Research on Rural Education: r^2ed.unl.edu

Sheridan, S. M., & Kratochwill, T. R. (2008). *Conjoint behavioral consultation*. New York: Springer.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*, 1-22. doi: 10.1037/a0029312

Snijders, T. A. B., & Bosker, R. J. (2012*). Multilevel analysis: An introduction to basic and advanced multilevel modeling*. (2nd Ed.). London, UK: SAGE Publications, Ltd.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavioral Research, 44*, 1244-1254. doi: 10.3758/s13428-012-0213-1

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*, 325-346. doi:

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35*, 1-10.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Assessment and Intervention, 2*, 142-151. doi: 10.1080/17489530802505362.

White, G. C., & Bennetts, R. E. (1996). Analysis of frequency count data using the negative binomial distribution. *Ecology, 77*, 2549-2557.

Appendix A: Participants, Instruments, and Procedures

Table A1.

Externalizing and Internalizing Behaviors on the Parent Daily Report (PDR)

Aggressiveness	Defiance	Lying	Hyperactiveness	Running around	School contact
Destructiveness	Fire setting	Pouting	Irritableness	Running away	Parents spank
Bedwetting	Fearfulness	Soiling	Noncomplying	Sadness	Noisiness
Competitiveness	Whining	Stealing	Negativism	Temper tantrum	Police contact
Complaining	Arguing	Yelling	Not eating meals	Talking back	
Hitting others	Crying	Teasing	Pants wetting	Fighting	

Appendix B: Results for the Theoretical Multilevel Model

Table B1.

Summary of Covariance Parameter Estimates for Theoretical Model

	Covariance Parameter	Estimate	SE
Normal (ALL DDFM)	Intercept (C[T])	17.85	2.05
	Phase (C[T])	3.51	.80
	Intercept (T)	-1.56	1.31
	Phase (T)	-0.49	.48
	Residual	10.94	.30
Normal + Sandwich Estimator (BW, CON, RES)	Intercept (C[T])	17.85	2.05
	Phase (C[T])	3.51	.80
	Intercept (T)	-1.56	1.13
	Phase (T)	-0.49	.48
	Residual	10.94	.30
Poisson (ALL DDFM)	Intercept (C[T])	.41	.048
	Phase (C[T])	.19	.027
	Intercept (T)	-.016	.032
	Phase (T)	-.027	.015
Negative Binomial (ALL DDFM)	Intercept (C[T])	.40	.049
	Phase (C[T])	.16	.027
	Intercept (T)	-.015	.033
	Phase (T)	-.027	.015
	Residual	.093	.007

Note: Intercept and phase effects for children (C) nested within teachers (T). Kenward Roger and Satterthwaite are unavailable for normal distribution with sandwich estimator.

Table B2.

Summary of the Solutions for Fixed Effects for Theoretical Model

	Effect	Est	SE	Sign.
Normal	Intercept	7.99	.215	< .0001
	Phase	-2.11	.143	< .0001
Normal + Sandwich Estimator (BW, CON, RES)	Intercept	7.99	.231	< .0001
	Phase	-2.11	.154	< .0001
Poisson	Intercept	1.92	.034	< .0001
	Phase	-.387	.026	< .0001
Negative Binomial	Intercept	1.92	.034	< .0001
	Phase	-.382	.025	< .0001

Note: Significance for containment is unavailable for normal, Poisson, and negative binomial distributions (Den DF = 0). Kenward Roger and Satterthwaite degree of freedom methods are unavailable for normal distribution with sandwich estimator (default = residual).

Table B3.

Stage 1 Population Parameter Estimates to Inform Stage 2 Data Generation

Poisson		**Negative Binomial** [2]	
Level-two intercept	π_{0jk} =.4136	Level-two intercept	π_{0jk} =.4031
Level-two phase effect	π_{1jk} = .1896	Level-two phase effect	π_{1jk} = .1558
Level-three intercept[1]	β_{00} = .05786	Level-three intercept[1]	β_{00} = .05744
Level-three phase effect[1]	β_{01} = .02652	Level-three phase effect[1]	β_{01} = .02461
Fixed intercept	γ_{000} = 1.9199	Fixed intercept	γ_{000} = 1.9172
Fixed phase effect	γ_{100} = .3865	Fixed phase effect	γ_{100} = .3818

Note: Based on the calculated unconditional ICC.[1]
Data generated as Negative Binomial was simulated using a probability of .502. [2]

Appendix C: Results for the Data-Driven Multilevel Model

Table C1.

Covariance Parameter Estimates and Solutions for Fixed Effects of the Data-Driven Model with Normal Distribution

	Empty Model					
Parameter	Estimate	SE	DF	t value	Sig.	
Intercept (γ_{00})	6.7191	.09069	3452	74.09	< .0001	
Scale (ϕ)	28.3989	.6835	-	-	-	
	Random Intercept at Level 2			**Random Intercept at Level 2 with Fixed Effect**		
Covariance Parameter	Estimate	SE		Estimate	SE	
Intercept (τ_0^2) [ChildID(TeacherID)]	15.6627	1.2623		15.7816	1.2594	
Residual (σ^2)	13.0219	.3326		11.8396	.3024	
Fixed Effects Solution	Estimate	SE	Sig.	Estimate	SE	Sig.
Intercept (γ_{00})	6.7837	.2129	< .0001	8.0016	.2238	< .0001
Phase (γ_{01})	-	-		-2.1198	.1209	< .0001
Type III Tests of Fixed Effects				Num DF, Den DF	F Value	Sig.
Phase	-	-		1, 3071	307.34	< .0001

Note: *Normal distribution.*

Table C2.

Covariance Parameter Estimates and Solutions for Fixed Effects of Data-Driven Model with Poisson Distribution

	Empty Model				
Parameter	Estimate	SE	DF	t value	Sig.
Intercept (γ_{00})	1.905	.006565	3452	290.16	< .0001
Scale (ϕ)					

	Random Intercept at Level 2			Random Intercept at Level 2 with Fixed Effect		
Covariance Parameter	Estimate	SE		Estimate	SE	
Intercept (τ_0^2) [ChildID(TeacherID)]	.4646	.03736		.4632	.03729	
Fixed Effects Solution	Estimate	SE	Sig.	Estimate	SE	Sig.
Intercept (γ_{00})	1.7135	.03599	< .0001	1.8791	.03655	< .0001
Phase (γ_{01})	-	-	-	-.3091	.01335	< .0001
Type III Tests of Fixed Effects				Num DF, Den DF	F Value	Sig.
Phase	-	-		1, 3071	536.13	< .0001

Note: *Poisson distribution.*

Table C3.

Covariance Parameter Estimates and Solutions for Fixed Effects for Data-Driven Model with Negative Binomial Distribution

	Empty Model					
Parameter	Estimate	SE	DF	t value	Sig.	
Intercept (γ_{00})	1.905	.01446	3452	131.74	< .0001	
Scale (ϕ)	.5732	.01836	-	-	-	
	Random Intercept at Level 2			**Random Intercept at Level 2 with Fixed Effect**		
Covariance Parameter	Estimate	SE		Estimate	SE	
Intercept (τ_0^2) [ChildID(TeacherID)]	.4417	.03717		.4515	.03761	
Scale	.1626	.00938		.13	.008278	
Fixed Effects Solution	Estimate	SE	Sig.	Estimate	SE	Sig.
Intercept (γ_{00})	1.7721	.03588	< .0001	1.9056	.03752	< .0001
Phase (γ_{01})	-	-	-	-.3471	.01935	< .0001
Type III Tests of Fixed Effects				Num DF, Den DF	F Value	Sig.
Phase	-	-		1, 3071	321.94	< .0001

Note: Negative Binomial distribution

Table C4.

Log Likelihood Difference Test for Data-Driven Model with Normal, Poisson, and Negative Binomial Distributions

		-2LL (full)	-2LL (nested)	Sig.
Normal	Empty Model	21354.14		
	Random intercept at level 2	21354.14	19592.1	$p < .0001$
	Random intercept at level 2 w/ fixed phase effect	19592.1	19299.1	$p > .0001$
Poisson	Empty Model	26452.26		
	Random intercept at level 2	26452.26	19547.21	$p < .001$
	Random intercept at level 2 w/ fixed phase effect	19547.21	19015.74	$p > .001$
Negative Binomial	Empty Model	20245.44		
	Random intercept at level 2	20245.44	18709.12	$p < .001$
	Random intercept at level 2 w/ fixed phase effect	18709.12	18398.37	$p > .001$